STANDARD SETTING FOR SPECIALIST PHYSICIAN EXAMINATIONS IN SOUTH AFRICA

by

DR FRANS HENDRIK SCARPA SCHOEMAN

Thesis submitted in fulfilment of the requirements for the degree Philosophiae Doctor in Health Professions Education

Ph.D. HPE

in the

DIVISION HEALTH SCIENCES EDUCATION

FACULTY OF HEALTH SCIENCES

UNIVERSITY OF THE FREE STATE

BLOEMFONTEIN

JANUARY 2015

INTERNAL PROMOTER: **PROF. DR M.M. NEL**

EXTERNAL PROMOTER: **PROF. DR V.C. BURCH**

DECLARATION

I hereby declare that the work submitted here is the result of my own independent investigation. Where help was sought, it was acknowledged. I further declare that this work is submitted for the first time at this university/faculty towards a Philosophiae Doctor degree in Health Professions Education and that it has never been submitted to any other university/faculty for the purpose of obtaining a degree.

…… January 2015

F.H.S SCHOEMAN Date

I hereby cede copyright of this product in favour of the University of the Free State.

…… January 2015

F.H.S SCHOEMAN Date

DEDICATION

I would like to dedicate this thesis to my wife (Marlene) and sons (Christian and Daniel) - your love, support, patience and faith in me have made this dream possible.

I also dedicate this work to all the medical educators and trainees in South Africa, as well as the patients we serve. I hope this thesis will contribute to better health and career progression outcomes for all involved.

"Test difficulty is like love, it's in the eye of the beholder" – Scarpa Schoeman

ACKNOWLEDGEMENTS

I would like to thank everyone who helped to make this thesis possible. I wish to express my sincere thanks and appreciation to the following persons and organisations:

- My internal promoter, Prof. Marietjie Nel, Former Head: Division of Health Sciences Education, Faculty of Health Sciences, University of the Free State, for her support, encouragement and supervision. Thank you for assisting me and believing in me to get over the finishing line.

- My external promoter, Prof Vanessa Burch, Chair of Clinical Medicine, Department of Medicine, University of Cape Town, for her support and leadership, guidance, supervision and nights spent reading my chapters. Your faith in me meant a lot and huge thanks for championing the change process in the College of Physicians of South Africa and this study.

- The National Research Foundation (NRF) Thuthuka programme, the Medical Research Council (MRC) and the Faculty of Health Sciences, School of Medicine, University of the Free State for the financial support provided to enable this project and disseminate its outcomes.

- The Council and examiners of the College of Physicians of South Africa, as well as the CMSA administration staff (Mrs Vorster and Botha in particular) - without your support, time, cooperation and valuable contributions, this project would not have been possible.

- Mrs Yolanda Nagel, my assistant at the Department of Internal Medicine, University of the Free State, for supporting me in this project and carrying the administrative load when I was away on study leave. You are a pillar of strength and reliability.

- Ms Nadia Laubscher and Melody Mentz, statisticians of MelodyM Consulting, and Prof Gina Joubert, Department of Biostatistics, University of the Free State for help and advice with some of the statistical data analysis of the study.

- Prof Vanessa Burch and Mrs Dot Bransby for the final language editing of the thesis.

- Dave and Beverley Bell for allowing me to use their SurveyMonkey® software. I really appreciate your help and support in this regard tremendously.

- Dr Mark Allen of Wits University, for helping me to sort out certain aspects of the MCQ examinations data. Huge thanks to you for your time and effort. Much appreciated.

- My wife and sons, parents, parents-in-law, and other close family members. Many thanks for your support, encouragement and faith in me during the last three years.

- All my colleagues at the Department of Internal Medicine, University of the Free State, especially Drs Madelein Koning, Hannes Coetser and Prof Vernon Louw. Thank you for your support, encouragement and assistance.

- Most importantly, my HEAVENLY FATHER, without the strength YOU gave me this project would have been impossible.

TABLE OF CONTENTS

CHAPTER 1: ORIENTATION TO THE STUDY

CHAPTER 3: RESEARCH DESIGN AND METHODOLOGY

CHAPTER 5: FCP (SA) PART I MULTIPLE CHOICE QUESTION (MCQ) TEST COMPARING THE PERFORMANCE OF THE ANGOFF AND COHEN METHODS

CHAPTER 6: FCP (SA) PART II OBJECTIVE TEST (OT) COMPARING THE PERFORMANCE OF THE ANGOFF AND COHEN METHODS

APPENDIX E:
ITEM QUALITY INDEX PLOT EXPLANATION

APPENDIX F:
ETHICAL APPROVAL AND PERMISSION TO CONDUCT THE STUDY

LIST OF FIGURES

LIST OF TABLES

GLOSSARY OF TERMS USED

ABIM	**American Board of Internal Medicine**
cf.	***confer* (refer to)**
CHE	**Council on Higher Education**
CIV	**Construct-irrelevant variance**
CMSA	**Colleges of Medicine of South Africa**
CoP	**College of Physicians of South Africa**
DBE	**Department of Basic Education**
DI	**Discrimination Index**
DoHET	**Department of Higher Education and Training**
ECC	**Examinations and Credentials Committee (of the CMSA)**
EMI	**Extended Matching Items**
et al.	***et alii* (and others)**
etc.	***et cetera* (and so forth)**
FCP (SA)	**Fellowship of the College of Physicians of South Africa**
HCW	**Healthcare Worker**
HEQC	**Higher Education Qualifications Committee**
HPCSA	**Health Professions Council of South Africa**
HSE	**Health Sciences Education**
ID	**Item Difficulty**

i.e.	***id est* (that is)**
IQI	**Item Quality Index**
MCQ	**Multiple Choice Question**
MDPB	**Medical and Dental Professional Board**
MRCP	**Member of the Royal College of Physicians**
NBME	**National Board of Medical Examiners**
NQF	**National Qualifications Framework**
OSCE	**Objectively Structured Clinical Examination**
OT	**Objective Test**
PBTC	**Panel-based Test-centred**
PC-value	**Proportion Correct value**
RSA	**Republic of South Africa**
SA	**Social Accountability**
SAQ	**Short Answer Question**
SAQA	**South African Qualifications Authority**
SBA	**Single Best Answer**
SD	**Standard Deviation**
SEM	**Standard Error of Measurement**
SEQ	**Short Essay Question**
UFS	**University of the Free State**

UK	**United Kingdom**
USA	**United States of America**
WHO	**World Health Organisation**

SUMMARY

Key terms: **Angoff method, Assessment, Change management, Cohen method, Licensing examinations, Medical specialist certification, Postgraduate medical education, Quality assurance, Resource-limited assessment, Standard setting, Written assessment**

Setting defensible and fair pass standards for high-stakes postgraduate specialist certification examinations is a critical quality assurance component of assessment. Doing so in a feasible and sustainable way, within a resource-constrained context such as South Africa, is challenging.

Traditionally the 28 member Colleges of the Colleges of Medicine of South Africa (CMSA), the national specialist licensing examination body in South Africa, have used a fixed pass mark of 50%. This practice does not acknowledge the inherent variance in examination difficulty and so increases the risk of failing competent candidates (false negative outcome) and passing incompetent examinees (false positive outcome). In 2011, the College of Physicians (CoP), a large CMSA member College, addressed the matter by implementing a standard setting process for the written components of their specialist physician certification examinations.

The aim of this study was twofold: i) To evaluate the knowledge, attitudes, views and perspectives of CoP examiners regarding standard setting, and ii) compare the performance and utility of the Cohen and Angoff methods to advise the CoP regarding an appropriate standard setting method in a resource-constrained setting.

A literature review was done to conceptualise standard setting as it pertains to assessment in medical education. In addition, policies and regulatory systems relevant to specialist certification examinations in South Africa were reviewed to provide the context for this study.

Two research components were concurrently conducted between 2012 - 2014:
A prospective study evaluated the knowledge, attitudes, views and perspectives of CoP examiners regarding standard setting before and after training and 30 months

of practical experience using both the Cohen and Angoff methods of standard setting.

A comparative study evaluated the performance (pass marks and failure rates) and utility (according to a framework derived from the literature review) of the Cohen and Angoff methods using five cycles of examination data, including multiple choice questions (MCQ), short answer questions and short essay questions.

The introduction of standard setting was successful and widely supported by the CoP examiners. The Cohen method performed well when used for test data with a reasonable number of test items (30 or more) in homogeneous exit-level cohorts of more than 50 candidates. Tests containing few test items (i.e. short essay questions) performed poorly. The performance of the Cohen method was variable for smaller cohorts (less than 100) of candidates drawn from heterogeneous populations, such as entry-level Part I MCQ test takers. The Angoff method yielded unacceptable outcomes regardless of test format. The utility comparison identified the Cohen method as the preferred standard setting method for the CoP.

The findings of this study support the introduction and ongoing use of the Cohen method as a feasible and sustainable method of setting pass marks for the written components of the CoP certification examinations. Education and training in the use of standard setting methods, as part of a change management strategy, improved examiners' understanding of the role, importance and basic methodology of standard setting and strengthened their support for the use of standard setting in certification examinations. More data are needed to evaluate the true impact of cohort size on the stability of the Cohen method for entry-level, heterogeneous cohorts of examinees. The purist Angoff strategy, used in this study due to resource limitations, performed poorly and was deemed 'not fit for purpose' by the CoP examiners. The usefulness of the novel standard setting utility framework developed in this study warrants further research in other examination settings such as performance–based examinations.

OPSOMMING

Sleutelterme: **Angoff-metode, Assessering, Veranderingsbestuur, Cohen-metode, Lisensiëringseksamens, Mediese spesialis-sertifisering, Nagraadse mediese opvoedkunde, Gehalteversekering, Hulpbronbeperkte assessering, Slaagstandaard-bepaling, Skriftelike assessering**

Om verdedigbare en regverdige slaagstandaarde te bepaal vir nagraadse spesialis-sertifiseringseksamens, wat belangrike gevolge het, is 'n kritiese deel van gehalteversekering van assessering. Om dit te doen op 'n haalbare en volhoubare manier, in 'n hulpbronbeperkte konteks soos Suid-Afrika, is 'n uitdaging.

Tradisioneel het die 28 lid-kolleges van die Kolleges van Geneeskunde van Suid-Afrika (KGSA), die nasionale spesialis-lisensiëringseksamenliggaam in Suid-Afrika, 'n vaste slaagsyfer van 50% gebruik. Hierdie praktyk reflekteer nie die inherente variasie in eksamenmoeilikheidsgraad nie en verhoog so die risiko om bevoegde kandidate verkeerdelik te druip (vals-negatiewe uitkoms) en om onbevoegde kandidate te laat slaag (vals-positiewe uitkoms). In 2011 het die Kollege van Interniste (KvI), 'n groot KGSA lid-kollege, hierdie situasie aangespreek deur 'n proses te implementeer om die slaagstandaard (slaagsyfer) van die skriftelike komponente van hulle spesialis-internis sertifiseringseksamens te bepaal.

Die doel van hierdie studie was tweeledig: i) evalueer die kennis, houding, sienings en perspektiewe van KvI eksaminatore ten opsigte van slaagstandaard-bepaling, en ii) vergelyk die prestasie en nuttigheid van die Cohen- en Angoff-metodes om sodoende die KvI te adviseer ten opsigte van 'n toepaslike metode vir slaagstandaard-bepaling, in 'n hulpbronbeperkte omgewing.

'n Literatuuroorsig is gedoen om slaagstandaard-bepaling te konseptualiseer, soos dit van toepassing is op assessering in mediese opvoedkunde. Daarbenewens, is beleide en regulatoriese sisteme, relevant tot spesialis-internis sertifiseringseksamens in Suid-Afrika, ook hersien en bygewerk om die konteks van die studie te skets.

Twee navorsingskomponente was gelyktydig uitgevoer tussen 2012 – 2014:

'n Prospektiewe studie het die kennis, houdings, sienings en perspektiewe van KvI eksaminatore geëvalueer ten opsigte van slaagstandaard-bepaling, voor en na opleiding en 30 maande se praktiese ervaring in die gebruik van beide die Cohen- en Angoff-metodes van slaagstandaard-bepaling.

'n Vergelykende studie het die prestasie (slaagsyfers en druipkoerse) en nuttigheid ('n raamwerk wat uit die literatuuroorsig ontwikkel is) van die Cohen- en Angoff-metodes ondersoek deur vyf siklusse van eksamendata te evalueer, insluitende veelvuldige keusevrae (VKV), kort-antwoordvrae en kort-opstelvrae.

Die ingebruikneming van slaagstandaard-bepaling was suksesvol en het wye steun geniet onder die KvI eksaminatore. Die Cohen-metode het goed presteer op toetsdata met 'n redelike aantal toetsitems (30 of meer) in homogene, uittreevlak kohorte van 50 kandidate of meer. Toetse met min items (kort-opstelvrae) het swak vertoon. Die prestasie van die Cohen-metode was variërend vir kleiner kohorte (minder as 100) van kandidate uit heterogene populasies, soos die intreevlak Deel I VKV toetsnemers. Die Angoff-metode het onaanvaarbare uitkomste gelewer, ongeag die toetsformaat. Die nuttigheidsvergelyking het die Cohen-metode geïdentifiseer as die voorkeurmetode vir slaagstandaard-bepaling in die KvI.

Die bevindinge van hierdie studie ondersteun die ingebruikneming en voortgesette gebruik van die Cohen-metode as 'n haalbare en volhoubare metode om die slaagsyfers van die skriftelike dele van die KvI se sertifiseringseksamens te bepaal. Onderrig en opleiding in die gebruik van slaagstandaard-bepalingsmetodes, as deel van 'n veranderingsbestuurstrategie, verbeter eksaminatore se begrip van die rol, belangrikheid en basiese metodiek van slaagstandaard-bepaling en versterk hulle steun vir die gebruik van slaagstandaard-bepaling in sertifiseringseksamens. Meer data is nodig om die ware impak van kohortgrootte op die stabiliteit van die Cohen-metode vir intreevlak-heterogene kohorte van kandidate te evalueer. Die puristiese Angoff-strategie, wat weens hulpbrontekorte in hierdie studie gebruik is, het swak presteer en is as 'nie-geskik vir die doel' deur die KvI eksaminatore geag. Die bruikbaarheid van die nuwe slaagstandaard-bepaling nuttigheidsraamwerk, wat ontwikkel is in hierdie studie, regverdig verdere navorsing in ander eksamenomgewings, soos prestasie-gebaseerde eksamens.

STANDARD SETTING FOR SPECIALIST PHYSICIAN EXAMINATIONS IN SOUTH AFRICA

CHAPTER 1

ORIENTATION TO THE STUDY

1.1 INTRODUCTION

The aim of this chapter is threefold. Firstly, it is to orientate the reader to this Ph.D. study and thesis outline. Secondly, to introduce the general concepts and principles of standard setting as it relates to assessment in health sciences education. Thirdly, it aims to highlight the importance of standard setting in the quality assurance process as it pertains to assessment in health sciences education.

1.1.1 Orientation to the study

In this Ph.D. research project, the researcher investigated three aspects of standard setting.

Firstly was the introduction of a formal standard setting process for written specialist licensing examinations in one of the largest colleges (national specialist licensing examination body) in the Colleges of Medicine of South Africa (CMSA), namely the College of Physicians of South Africa (CoP). As part of this process of introducing standard setting, the knowledge, attitudes, views and perspectives on standard setting of the current examiners in the CoP were assessed before and after a training intervention and 30 months exposure of these examiners to two methods of standard setting.

Secondly, the study compared the quantitative performance (pass marks and failure rates) of two standard setting methods, the Angoff and Cohen methods, using five cycles of written CoP examinations over a three-year period.

Thirdly, the study evaluated the overall utility of the Cohen method in comparison to the Angoff method, in the context of the CoP, after 30 months of exposure to both

methods. This evaluation was based on utility parameters of standard setting methods, which were identified and derived from the literature review in Chapter 2.

The next section of this chapter provides an orientation to standard setting in health sciences education (section 1.1.2). The rest of the chapter describes the motivation to undertake this research project, the problem statements and research questions investigated, the overall goal, aim and objectives of the study, the demarcation of the field and scope of the study, the significance and value of the study, the research design and methods of the study, implementation of the findings and it ends with an overview of the lay-out of this Ph.D. thesis and a concluding summary.

1.1.2 Orientation to standard setting in health sciences education

An orientation to standard setting in health sciences education requires a closer look at assessment of student performance and the role standard setting plays in assuring the quality of assessment practices.

1.1.2.1 *Health sciences education, assessment and standard setting*

In health sciences education, faculty members (academic staff members in higher education and training institutions) and governmental regulatory bodies of specific professions such as medicine, nursing or physiotherapy, for example, must determine the standards that students in the respective disciplines need to attain in order to be licensed as competent to practise their respective profession. Since the concept of educational 'standards' is abstract or conceptual (Barman 2008:957), it needs to be clearly defined and articulated to enable academic staff members and other stakeholders to agree on standards and subsequently implement them. The process of converting abstract educational standards into something concrete and measurable is called curriculum development and it leads to the production of programme (or curricular) outcomes (Grant 2010:2-3). It is against these outcomes that student performances are measured, through the assessment system, to determine whether they have achieved the programme outcomes at a satisfactory level, a given standard, and are therefore deemed competent to progress or not (Grant 2010:14).

Since it is not possible to test student achievement for all programme outcomes, a sample of outcomes is assessed at various intervals of the programme using

different types of assessment instruments. The results of these assessment 'biopsies' serve as samples of the entire population of programme outcomes and are used to provide feedback to the students on their progress towards mastery of the programme outcomes (formative assessment) (Norcini, Anderson, Bollela, Burch, Costa, Duvivier, Galbraith, Hays, Kent, Perrott & Roberts 2011:211; Schuwirth & Van der Vleuten 2010:195; Wood 2010:259), or to make judgement decisions as to which students have mastered the outcomes sufficiently to progress to the next level or graduate from the programme as competent practitioners (summative assessment) (Hill, Guinea & McCarthy 1994:394; Norcini *et al.* 2011:211; Schuwirth & Van der Vleuten 2010:195; Wood 2010:260).

It is apparent from the preceding discussion that there are many potential areas where assessment processes face significant challenges in both undergraduate and postgraduate programmes. These include, amongst others:

- Not all training standards are converted to outcomes and assessed, leaving the student with competency 'gaps' after completing their training, which will only be exposed once they start to practise.
- Some outcomes may never have been addressed or taught in the programme, but may be included in the assessments, the so-called 'hidden curriculum' (Shumway & Harden 2003:574; Van der Vleuten 1996:51).
- Some assessments may include a disproportionate number of outcomes from a particular segment of the programme and hence will render that assessment unrepresentative of the entire population of outcomes. This leads to a false or inaccurate deduction of the students' mastery of the programme outcomes (Schuwirth & Van der Vleuten 2010:196).
- Instruments used to assess mastery of programme outcomes may be of a poor quality, leading to results which are unreliable (Schuwirth & Van der Vleuten 2010:196).

Assessment strategies include, to a greater or lesser extent, elements of all of the above-mentioned errors. For this reason, regular feedback from stakeholders, including students, is needed to identify problem areas in the programme, particularly with regards to assessment. Such feedback allows educators and examiners to address the identified areas of concern (Grant 2010:4).

Although assessment systems might be fraught with potential challenges as described above, the critical question remains: How many programme outcomes should a student have mastered, as measured by the assessment system, in order to be deemed competent to practise as a healthcare practitioner or specialist practitioner in the case of postgraduate education? Realistically, no student will master all of the outcomes perfectly, and conversely, all students will master most of the outcomes to a greater or lesser extent. So, where on this continuum of student performance is the point or score which determines whether a student passes or fails the test or examination? The answer to this question forms the basis for all standard setting activities and is at the heart of summative assessment. It has been described by Schuwirth and Van der Vleuten (2010:204) as the 'Holy Grail' (highest or most sought after goal) of assessment in health sciences education.

1.1.2.2 *Background to the concept and importance of standard setting*

Standard setting has been discussed in the literature for more than 50 years and many methods of setting standards have been described and proposed (Barman 2008:958; Cusimano 1996:112; Downing, Tekian & Yudkowsky 2006:50). In broad terms, formal standard setting methods and processes have been developed to help educators determine which candidates, sitting for a particular test or examination, have performed well enough to pass the assessment and which have not. The issue of what constitutes the minimum acceptable level of competence to pass an assessment has been of particular interest and concern to educators in the health sciences, because the stakes of these decisions are high for all stakeholders involved.

Boelen (1995:S26) has stressed that health sciences education institutions and educators are also responsible for setting minimum levels of competence (standards) to ensure patient safety (social responsibility) and are also accountable for ensuring that these standards are maintained (social accountability). Therefore, to enable institutions and educators to live up to these societal demands and expectations, effective standard setting procedures need to be in place, together with other sound teaching, learning and assessment policies, procedures and practises (Lindgren & Karle 2011:668). A broader discussion on the social accountability of standard setting is offered in Chapter 2.

Occasionally standard setting processes lead to the unfortunate misclassification of students or trainees in terms of passing and failing high-stakes examinations. If a competent student fails during the final licensing examination (a false negative result), the candidate suffers significant emotional distress, disappointment and possibly incurs an additional financial burden, including the potential loss of a scholarship or bursary (Bandaranayake 2008:837; Cusimano 1996:117). In contrast to this, and possibly of more concern, is a false positive test result in which an incompetent candidate passes the examination. This outcome allows incompetent practitioners to practise, putting patients and employers at risk (Bandaranayake 2008:837). Furthermore, this outcome also negatively impacts on the trust the public, university and government have in the academics and administrators who conduct high-stakes examinations.

It is, therefore, important that assessment systems in health sciences education should not only be fair to students in terms of what is assessed in the test or examination (content), but also be of the highest possible quality (Schuwirth & Van der Vleuten 2010:198) in terms of validity, reliability and administration (process). Standard setting is part of the process of administering assessments and making pass/fail decisions which are credible and defensible (Bowers & Shindoll 1989:1). Therefore, standard setting plays a fundamental role in determining the quality of assessment systems. Without standard setting, assessment systems may produce increased false positive or false negative test results, which are detrimental to both candidates and patients seeking healthcare (Bandaranayake 2008:837; Cusimano 1996:117). This undermines the social accountability mandate of health science institutions and their training programmes.

While standard setting is applicable to *all disciplines* in health sciences education and indeed to all aspects of higher education, this thesis will focus on quantitative standard setting (using marks or scores), as it pertains to the *written licensing examinations for specialist physicians in South Africa.* Henceforth, the term 'medical education' will be used, since this thesis is restricted to the training of specialist physicians. The *qualitative* nature of some forms of standard setting in criterion-referenced assessment, using narrative grading criteria and feedback, in higher education is well-recognised (Sadler 2005:179-186), but falls outside the scope of this thesis and will therefore not be discussed further in any depth.

1.1.2.3 *Principles of standard setting*

As mentioned in the preceding text, standard setting is a cornerstone of the quality assurance process of an assessment system (Norcini *et al.* 2011:211; Schuwirth & Van der Vleuten 2010:204). There are some important principles and concepts that are relevant to the process of setting the pass mark or pass standard for assessments in medical education. The key principles are briefly discussed here.

- *A 'GOLD standard' does not exist*

  There is no 'gold standard' methodology of setting pass standards, which is suitable for all types of assessment strategies used (Barman 2008:958; Cohen-Schotanus & Van der Vleuten 2010:154; Cusimano 1996:112; Downing *et al.* 2006:51; Schuwirth & Van der Vleuten 2010:204). As a consequence, a variety of methods are in use which were designed to be used with specific assessment instruments (Schuwirth & Van der Vleuten 2010:205). Examples include the Angoff method for any written assessment (Angoff 1971), Nedelsky method for MCQs (Nedelsky 1954) and the Borderline Regression method (Wood, Humphrey-Murto & Norman 2006) for Objectively Structured Clinical Examinations (OSCEs).

- *Based on human judgement*

  All methods of standard setting involve human judgement (Bandaranayake 2008:837; Barman 2008:959; Crocker & Zieky 1995:26; Cusimano 1996:112; Downing *et al.* 2006:51; Norcini 2003:464). The way in which these judgments are made depends on the method used. Most of the established methods rely on a panel of expert judges who decide how a hypothetical borderline (i.e. a minimally- or just-competent) candidate would fare on the individual questions or items of a test. This strategy potentially introduces an amount of bias as each judge has his or her own idea about what constitutes borderline performance. In contrast, other methods such as the Cohen (Cohen-Schotanus & Van der Vleuten 2010:157) or Hofstee (Bandaranayake 2008:840; Downing *et al.* 2006:54; Hofstee 1983) methods use panels of judges to make decisions about the application of a flexible model to a set of test results, to determine the pass mark.

- *Standards must be explicable, defensible and stable*

  As just discussed, all pass standards are opinion-based and therefore, somewhat arbitrary in nature (Crocker & Zieky 1995:26; Kane 1994:426). However, they are

never without real importance and, as such, they must be explicable, defensible and stable among cohorts of students (Schuwirth & Van der Vleuten 2010:205). The key difficulty in dealing with pass standards arises when an institution or examination body *cannot* answer the following questions as discussed below:

a) Is the particular pass mark valid and credible? If no formal process or method was used or the method used does not attempt to link the difficulty of the assessment with the abilities of the candidates sitting the assessment, then the pass mark is deemed 'not defensible';
b) How does the method work which was used to determine the pass mark? If the method used is too complex for lay people or stakeholders to understand it, the pass mark may be deemed 'not explicable';
c) Why does the pass mark (or cut-score) vary so considerably between different cohorts of students? A wide variation in the pass mark between consecutive sittings of the same assessment, with similar cohorts of candidates, equivalent test content and conditions renders the pass mark 'not stable' (Schuwirth & Van der Vleuten 2010:205).

- *Processes and procedures must be feasible and sustainable*
  Whichever method(s) of standard setting are chosen by a particular institution for its assessment instruments, they must be feasible and sustainable (Barman 2008:960). Standard setting methods which are expensive in terms of time and human resources may not be acceptable and feasibility studies need to be done before a particular method is adopted for routine use (Berk 1986:143).

1.1.2.4 *Types of standard setting methods (classification)*

As mentioned in the previous sections, there are many different kinds of standard setting methods. In this sub-section, a brief discussion of the classification of the different standard setting methods is provided. Chapter 2 provides a more detailed discussion on the methods and concepts described here.

Absolute or criterion-referenced methods

Test-centred methods

These are methods that are based on the collective decision made by a panel of expert judges who set the pass standard, usually prospectively, for a particular test or

examination. These methods hinge on two concepts. First, is the concept of what constitutes a hypothetical 'borderline' candidate for the cohort in question (Norcini 2003:465). This is defined as the particular abilities and characteristics of a candidate sitting the assessment, who is minimally competent or just competent. Second, the panel of judges must use the first concept mentioned and use it to predict how this hypothetical borderline candidate *would* perform on each individual test item (Downing *et al.* 2006:53). The panellists' ratings or judgements are then summed and averaged across the test items and all the judges to produce the collective mean score for a just competent (borderline) candidate on the test (Norcini 2003:465). This is the minimum score (pass mark) that candidates need to achieve in the test or examination, in order to pass.

These methods provide an absolute pass mark, based on the perceived collective difficulty of the individual test items in the eyes of expert judges (Barman 2008:958; Cusimano 1996:113; Downing *et al.* 2006:51; Norcini 2003:465; Schuwirth & Van der Vleuten 2010:204). Hence, all the candidates who achieve a score at or above this absolute/criterion referenced pass mark are classified as competent to pass. These methods are deemed most appropriate for high-stakes, competency-based assessments, such as medical licensing examinations. Commonly used examples include the Angoff and Ebel methods (Bandaranayake 2008:839; Barman 2008:959; Cusimano 1996:113; Norcini 2003:465).

<u>Examinee-centred methods</u>

These methods also produce absolute, criterion-referenced pass marks, but in a different way than test-based methods. They are based on the judgements of a group of expert judges, who observe and rate the performance of candidates in performance-based assessments, such as OSCEs (McKinley & Norcini 2014:99). Apart from assessing and grading the tasks in the station, the judges also need to identify the candidates who demonstrate overall 'borderline' performance (just competent or minimally competent performance) in the particular station they are assessing (global rating). The marks derived from grading the station's tasks and the overall or global rating of the candidates, in particular the borderline candidates, are then used in different mathematical ways to set the pass standard retrospectively for each station and collectively for the whole test or examination. It is important to note that the number of candidates sitting the performance assessment must be adequate for these methods to be reliable and defensible (McKinley & Norcini 2014:102). Examples of

these methods include the Borderline Regression method (Wood *et al.* 2006:115) and the Borderline Group method (Wilkinson, Newble & Frampton 2001:1044).

Relative or norm-referenced methods

These are methods where the pass mark is determined retrospectively once the assessment event has been completed by the candidates. In *pure* norm-referencing, the judges decide on the passing criteria in terms of what number or percentage of students should pass the test, for example the top 60 students or the top 50% of students. Once the test has been administered and the results are available, the criterion is applied and, depending on the relative performance of the group of examinees, the pass mark is determined (Downing *et al.* 2006:51). Norm-referenced methods are only appropriate if a limited number of passing candidates can be accommodated in the subsequent years of a programme, i.e. the selection of a limited number of candidates is the main purpose of the assessment process (Norcini 2003:464). It is important to note that the actual competence of the passing candidates is not established or measured (Barman 2008:958). The eventual pass mark is essentially determined by the academic strength of the cohort being assessed and the inherent difficulty of the assessment items. An example of such a method is the modified Wijnen method, which sets the pass mark at the mean of the test scores minus one standard deviation (Schuwirth & Van der Vleuten 2010:204).

Compromise methods

These are methods that aim to combine the strengths and minimise the weaknesses of both criterion-referenced and norm-referenced methods used to set the pass standard (Cohen-Schotanus & Van der Vleuten 2010:157). They contain some aspects of norm-referencing, mostly to gauge the difficulty of the test, but they set an absolute/criterion pass mark for the assessment. Compromise methods are increasingly being used to improve the acceptability of standard setting processes (Bandaranayake 2008:840). Examples of such methods are the Hofstee and Cohen methods (Bandaranayake 2008:840; Cohen-Schotanus & Van der Vleuten 2010:157).

1.2 BACKGROUND TO THE RESEARCH PROBLEM

1.2.1 Two examination processes used for specialist certification in South Africa

The Health Professions Council of South Africa (HPCSA) is the statutory body responsible for the accreditation and regulation of all aspects of medical education and training in South Africa. This organisation also administers the initial and annual registration of all doctors and other healthcare professionals practising in South Africa. Furthermore, postgraduate medical specialist training in South Africa only takes place at HPCSA-accredited clinical training centres, situated in the eight medical schools operating in South Africa.

Postgraduate medical specialist trainees, called registrars or residents, are employed for a period of up to five years, depending on the speciality, in registered training posts in clinical departments such as Internal Medicine or Surgery, etc. In addition to completing the minimum residency period in a training post, registrars must undertake a licensing examination to qualify as a medical specialist in South Africa. Historically, there were two separate parallel examination routes. One route was via the local specialist department's own in-house postgraduate examinations, with external examiners, and the other option was through the national Colleges of Medicine of South Africa (CMSA) examinations. The latter is referred to as 'Fellowship' examinations, offered by the respective specialist colleges of the CMSA.

The CMSA, established in 1955, is a national postgraduate examination and certification body, consisting of 28 constituent specialist colleges. The organisation administers certification examinations in all the registered medical and dental specialist fields in South Africa. The CMSA is the equivalent of the Royal Colleges (of medical and dental specialities) in the United Kingdom (UK), Canada and Australasia. Trainees who have completed the relevant college's examinations graduate as 'Fellows' and are eligible to register as specialists with the regulator, the HPCSA. In the case of Internal Medicine, it is called the Fellowship of the College of Physicians of South Africa - FCP (SA) examinations. More detail on the three written papers of the FCP (SA) examination is provided in the next section.

1.2.2 A single exit examination for specialist certification in South Africa

In 2010 the HPCSA decided that all exit licensing examinations for postgraduate specialist training in South Africa should be done through one central examining body (HPCSA 2010:2). This strategy was introduced as a means of centralising and standardising the exit examinations for all specialist trainees in South Africa, irrespective of where in the country they trained. Towards the end of 2010, the CMSA was mandated by the HPCSA to be the central examining body hosting all postgraduate specialist exit-level licensing examinations (HPCSA 2011). As a result, all doctors who commenced postgraduate specialist training as of 1 January 2011, are required to sit the CMSA exit licensing examinations (HPCSA 2011). Although the decision was implemented from 2011, the memorandum of understanding between the two parties was only finalised in June 2014 (HPCSA & CMSA 2014). Since the CMSA, for the foreseeable future, will be the only specialist licensing examinations body in South Africa, the quality and rigour of the CMSA examinations are of utmost importance.

The CoP oversees and administers the academic content of its FCP (SA) examinations. There are two parts to the FCP (SA) examinations – Part I (entry) and Part II (exit). Any doctor who has registered or is eligible to register with the HPCSA as having completed the mandatory two years of internship training, after qualifying as a doctor, can attempt the Part I examination (CoP 2013:2). The Part II examination is the exit licensing examination for trainees who have completed, or are nearing completion, of their four year residency training programme in Internal Medicine. In addition to having passed the FCP (SA) Part I examination and having spent enough time in training (a minimum of 2.75 years) in a registered training post, they also need to have their logbook of clinical experience signed-off by their supervisor to gain entry to the Part II examination (CoP 2013:2). A descriptive summary of Parts I and II of the FCP (SA) examination process, as it was during the period in which this study was conducted, is provided below.

<u>FCP (SA) Part I examination</u>

- The *Multiple Choice Question* (MCQ) written paper – a 150-item, single best answer (SBA) MCQ paper, with five possible options per item, aimed at assessing

the basic medical sciences (Physiology, Biochemistry, Pharmacology, etc.) as it pertains to Internal Medicine. Candidates have three hours to complete this test and generate a score out of a 150 marks (CoP 2013:4).

FCP (SA) Part II examination

- The *Objective Test* (OT) written paper - A constructed-response written paper consisting of 30 case-based questions of seven marks each, totalling 210 marks. It is focussed on clinical reasoning through assessing knowledge, understanding and insight into clinical case-scenarios, prompted by adjuncts such as x-rays, electrocardiograms, laboratory results and colour photographs of clinical conditions. Candidates have three hours to complete this test (CoP 2013:4).

- The *Short Essay Questions* (SEQ) written papers - Two mini-essay constructed-response written papers, each consisting of five questions of 30 marks each, usually broken down into two smaller sub-questions, totalling 150 marks per paper. Both papers focus on assessing theory knowledge, understanding and insight relating to the principles and practice of Internal Medicine. The two papers are written over two days and candidates have three hours to complete each paper. The marks from the papers are summed into a single score out of 300 marks. This combined mark is used to determine if a candidate has passed this component of the Part II written examination (CoP 2013:4).

- If a candidate passes both the OT and SEQ components *individually*, they are invited to the final component of the Part II examination – the bedside clinical examination. If a candidate passes this final clinical examination, s/he has then passed the FCP (SA) examination and is eligible to be admitted as a Fellow of the CoP. The clinical examination component of the Part II FCP (SA) examination falls outside the scope of this study and will not be discussed further.

The number of candidates annually sitting the FCP (SA) examinations over recent years (2010-2012) has been between 160-240 for Part I and 75-140 for Part II. However, given the 2010 HPCSA decision mentioned above and the CMSA mandate to conduct the exit examinations, these numbers are likely to increase over the coming years.

1.2.3 The need for standard setting as part of quality assurance

As discussed and explained in sections 1.1.2.1-2, determining the pass mark of an examination is an essential and crucial component of the quality assurance process of high-stakes assessment in medical education, particularly in postgraduate specialist certification examinations. Currently, a fixed examination pass mark of 50% is still used by all the constituent Colleges of the CMSA (CMSA 2015:online). This practice does not take into account the inherent variance in examination difficulty - which could potentially increase the false negative (failing the competent) or false positive (passing the incompetent) outcomes of the current examinations. In order to address this limitation, identification and implementation of one or more appropriate standard setting method(s), suitable for use in limited resource settings, would be an important step to enhance the quality of the certification examinations offered by the Colleges of the CMSA.

1.2.4 The introduction of standard setting and change management in the CoP

Since 2003, the CMSA has held several workshops and symposia focussing on assessment, with a view to promote best practice and evidence based practice in the assessment systems of its member Colleges (Burch 2014:1; Hift & Burch 2003:75-77). In response to the 2003 CMSA assessment symposium (Hift & Burch 2003:75-77), the CoP embarked on its own reform processes to improve the quality assurance (QA) processes of the FCP (SA) examinations of the CoP. These reforms were additionally driven by the published literature on assessment, exposure to expert colleagues in the field of medical education assessment as well as the research activities and publications of CoP council members and associates.

Figure 1.1 illustrates and highlights the significant assessment reforms that were introduced between 2002 and 2014 in the CoP. This information is provided to demonstrate that the CoP has been on a consistent road of assessment change and quality improvement since 2003, with some of the reforms published in the literature and discussed as part of the literature review of this thesis (Burch & Norman 2009:442-446; Burch, Norman, Schmidt & Van der Vleuten 2008:521-533). The present study is, therefore, a continuation of this assessment quality improvement and reform process in the CoP.

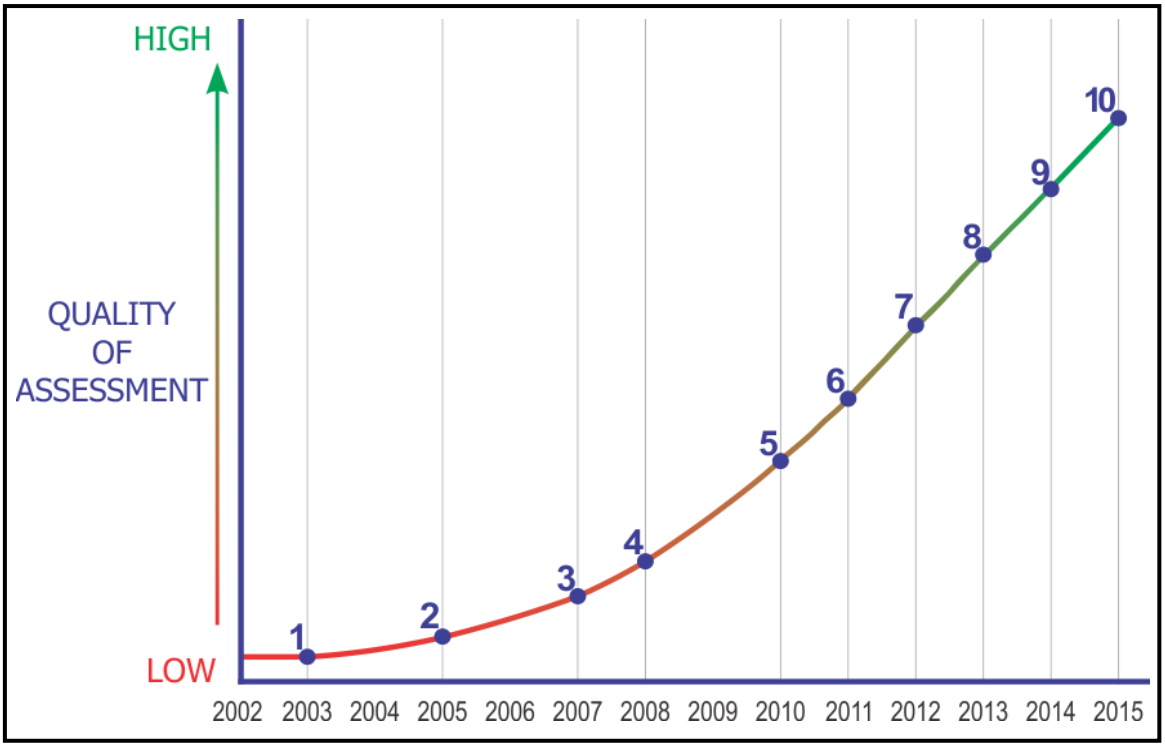


FIGURE 1.1: POSITIVE CHANGES TOWARDS INCREASING QA OF THE FCP EXAMINATIONS IN THE CoP
[Complied by the researcher, SCHOEMAN 2014]

Numbers 1- 10 below correspond and refer to Figure 1.1.

1) In May 2003, the first indication of a change process was noted as the CMSA held a two day symposium on the principles and quality assurance (QA) of postgraduate assessment (Hift & Burch 2003:75-77). The importance of standard setting, as part of the QA process of assessment, was noted in the report on the symposium by Hift and Burch (2003:75-77).

2-3) Between 2005 and 2007 several noteworthy reforms took place, culminating in the formalisation of the FCP (SA) examinations' syllabi, learning outcomes and study resources (Burch 2014:1; CoP 2011b:1-40).

4) Between 2007 – 2010 the research papers published, emanating from the Ph.D. research done by Vanessa Burch, a Council member and examiner of the CoP of the CMSA, which also called for reform and improved reliability in the FCP examinations, especially the final clinical examination (Burch 2007; Burch *et al.* 2008).

5) In 2010, the CoP changed the FCP Part I examination from a constructed-response, short essay question format to a 150-item MCQ test, blueprinted to the Part I syllabus (Burch 2014:1; CoP 2011b:4) .

During 2011, the CoP engaged in a formal process of implementing standard setting to its written FCP (SA) examinations. At the May 2011 CoP council meeting, the Cohen method of standard setting was selected (CoP 2011a:2) and gradually introduced over two years for the three written papers of the FCP (SA) examinations. The reasons why the Cohen method was selected are provided in the next section. The Cohen method implementation time frames were:

6) August 2011 for the Part I MCQ test (CoP 2011b:4);
7) March 2012 for the Part II OT (CoP 2013:4);
8) March 2013 for the Part II SEQ test (CoP 2014b:4).
9) Introduction of a single best answer from four possible options MCQ test as part of the FCP (SA) Part II (CoP 2014b:4).
10) Reform of the FCP (SA) clinical examination from 2015, as proposed by Burch *et al.* (2008:530).

One of the key challenges of implementing formal standard setting processes, in any setting, is a lack of knowledge about, and the limited use of standard setting processes by examiners and administrators of examining bodies. Since the CMSA is a well-respected and established examining body for postgraduate medical and dental specialist trainees, it was recognised that the introduction of standard setting procedures was likely to be a complex and sensitive matter that would need to be carefully managed to ensure that 'buy-in' and support for such a process was obtained and maintained without alienating examiners. This can only be achieved by employing a change management process, which:

- Adequately informs stakeholders about standard setting, thereby empowering them to make informed choices;
- Provides data, which enables stakeholders to make evidence-based decisions;
- Acknowledges existing resource limitations;
- Is informed by a clear understanding of the local context and organisational culture (Gale & Grant 1997).

While the use of formal standard setting procedures is strongly endorsed by the medical education literature, the knowledge, attitudes, views and perspectives of the CoP examiners regarding standard setting was not known at the time of

introducing the Cohen method. Hence, the potential need to provide training for the CoP examiners about standard setting and giving them 'hands-on' experience in the use of standard setting methods, as part of the change management process, was recognised and also investigated as part of this study.

1.2.5 Introduction of the Cohen method in the CoP

When the CoP decided to implement standard setting as part of their assessment practice, the choice of a specific method was largely guided by financial and human resource constraints (only 54 examiners and more than 200 candidates each year), which meant that a simple, user-friendly, resource efficient strategy was needed. Given the perceived simplicity and resource efficiency of the Cohen method, as compared to other methods like the Angoff method, it was selected for use, subject to review of its performance in the local context (CoP 2011a:2).

As mentioned, the CoP council wanted to formally evaluate and review the performance and utility of the Cohen method to a more widely used method, such as the Angoff method, in the CoP context, as part of the process of deciding about the long term use of the Cohen method (CoP 2011a:2). The process of evaluation was a catalyst for this study.

1.2.6 A brief description of the Cohen method

The Cohen method of standard setting, first published in the Dutch literature in 1996 (Cohen-Schotanus, Van der Vleuten & Bender 1996:83-87), and then in the English literature in 2010 (Cohen-Schotanus & Van der Vleuten 2010:157), holds much promise as a cost-effective and sustainable tool to determine the pass mark of summative examinations in a resource-limited setting such as South Africa - an explanation of why South Africa is a resource-limited setting is provided in Chapter 2.

In the Cohen method, the top-performing students are used as a point of reference to set an absolute pass mark. Essentially, the performance of the top candidates (90 - 95th percentile of the test scores) is used as the benchmark for the difficulty of the assessment and the pass mark is usually set at 60-70% of the benchmark (Cohen-Schotanus & Van der Vleuten 2010:159). The 95th percentile is

usually used because available research data suggests that this top cohort of candidates is quite stable (*cf.* Chapter 2, section 2.2.10) and performs equally well between different cohorts of examinees as compared to the mean test score, which is dragged down by poorly performing students (Cohen-Schotanus & Van der Vleuten 2010:157). In addition, the use of the 95th percentile also makes provision for 5% top-end outlier performers and hence, the outliers do not affect the pass standard for the cohort under assessment.

The actual numerator percentage used, to produce the pass standard from the benchmark percentile, is the collective policy agreement from a panel of judges. In this case, the CoP council decided on 65%. Therefore, for written assessments in the CoP, 65% of the 95th percentile represents the lowest acceptable score on the test that would be deemed as competent or a passing score. In this thesis, this particular application of the Cohen method is referred to as the 'Cohen65' pass mark. The Cohen method is illustrated graphically in Figure 1.2, as compiled by the researcher. This method will be discussed in greater detail in Chapter 2, section 2.2.10.

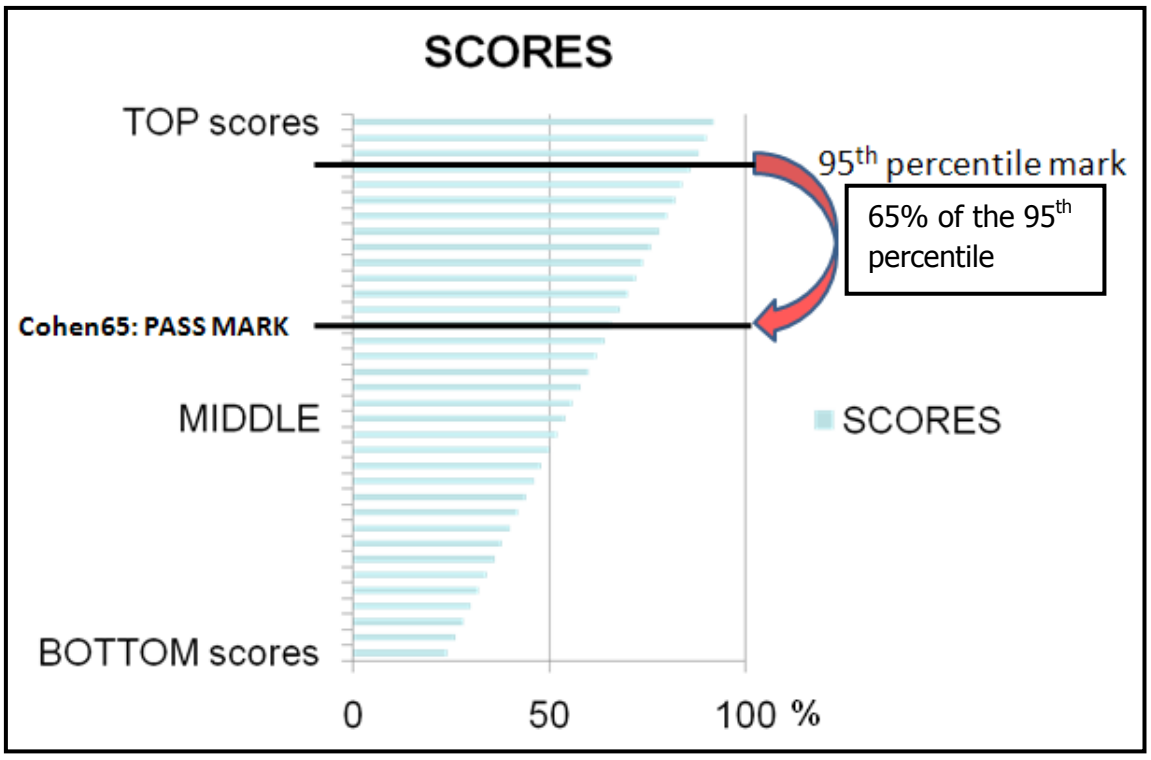


FIGURE 1.2: THE COHEN METHOD AS USED IN THE CoP
[Complied by the researcher, SCHOEMAN 2013]

1.2.7 The need for evaluation of the utility of the Cohen method

At the time of implementing the Cohen method of standard setting in the CoP, little was known in the published literature about the utility (as defined in Chapter 2) of this instrument, particularly in comparison with well recognised, widely used strategies, such as the Angoff method. Only one paper was found, which evaluated the Cohen method in comparison with the Angoff method on written assessments in medical education (Taylor 2011:e680). This study was done in one setting using test data of undergraduate medical students. The findings from this study are discussed further in Chapter 2. Given the limited amount of published data regarding the use of this method, further evaluation of the Cohen method, in particular its consistency (or stability) as compared to other established standard setting methods such as the Angoff method, was needed. Furthermore, at the time of initiating this study in 2012, there was no information about the use of the Cohen method in high-stakes postgraduate licensing examinations. For this reason, the CoP agreed to participate in a study which would provide data to determine the long term use of this method in the CoP context.

1.3 PROBLEM STATEMENT AND RESEARCH QUESTIONS

The problems or research topics that were identified during an extensive literature overview regarding the written assessment of postgraduate medical trainees gave rise to the research questions, which steered the process and methodology that was used in the study.

1.3.1 Problems identified

The problems identified and addressed by this study were as follows:

- The CoP was not using a recognised systematic method of determining the pass mark for the written assessments of the FCP (SA) examination. The traditional fixed 50% pass mark was in use.
- The knowledge, attitudes, views and perspectives of CoP examiners regarding standard setting were not known.

- The performance (pass marks and failure rates) and utility of the Cohen method, although a promising standard setting method, had not been evaluated in a range of assessment modalities used in postgraduate education and training context, such as those used in the FCP (SA) examination of the CoP.
- The Cohen method had not been subjected to evaluation of its utility, as compared to other established methods like the Angoff method, in postgraduate assessment systems.

As a result, this study was designed to:

- Determine the *knowledge, attitudes, views and perspectives* of CoP examiners with regard to standard setting, in order to provide appropriate training for examiners of the CoP.
- Determine and compare the *performance* (pass marks and failure rates) of the Cohen method, to that of the more established Angoff method, in setting pass standards for postgraduate written specialist examinations in the CoP.
- Determine and compare the *utility* of the Cohen method, to that of the more established Angoff method, as a standard setting method for routine use in postgraduate written examinations of the CoP.

1.3.2 Relevance of this study

This research project was relevant for two key reasons. Firstly, this research was an essential component of the process of introducing standard setting in the high-stakes written tests of the FCP (SA) examination. It was anticipated that this process would be challenging in the context of examiners who were not familiar with the concepts and principles of specific standard setting methods. Formal evaluation of the knowledge, attitudes, views and perspectives of examiners, with regard to standard setting in the CoP, facilitated the development of appropriate learning activities to provide a better understanding of standard setting and specifically, the Cohen method. Secondly, the performance and utility of the Cohen method, as a postgraduate standard setting method, required comprehensive evaluation.

1.3.3 Study hypotheses

This study's hypotheses were:

- Appropriate education and training in the use of standard setting methods, as part of a change management strategy, would help examiners develop a better understanding of the role, importance and basic methodology of standard setting and strengthen their support for the use of standard setting in the CoP.
- The Cohen method offers an acceptable way of setting pass standards for the high-stakes written examinations conducted by the CoP.
- The utility of the Cohen method is better than the Angoff method, in the context of the CoP.

1.3.4 Research questions

In order to address the aforementioned problems and to test the hypotheses, the following three research questions were derived for this study:

1. How is standard setting of assessment processes in medical education conceptualised in the literature and contextualised in postgraduate specialist certification examinations offered by the Colleges of Medicine of South Africa?

2. What are the knowledge, attitudes, views and perspectives of CoP examiners regarding standard setting, and do they change with training and exposure to a process of standard setting?

3. Is using the Cohen method, as compared to Angoff method, a more appropriate way of determining the pass mark for the written assessments of the FCP (SA) examination (Part I MCQ test, Part II Objective test and Part II SEQ test)?

1.4 OVERALL GOAL, AIM AND OBJECTIVES OF THE STUDY

In this study the goal denotes the broader vision that the researcher had in mind, while the aim refers to how it was envisaged to be achieved. The objectives are the specific steps that were taken to achieve the aim.

1.4.1 Overall goal of the study

The overall goal of the study was to improve the *quality* of the written assessment components of the FCP (SA) examination, by introducing a method of standard setting that is supported by local empiric research data, as gathered, analysed and presented in this thesis. In addition, this study may offer a way of improving the manner in which pass standards are currently set in other training programmes in health sciences education in South Africa.

1.4.2 Aim of the study

The aim of this study was to introduce and critically evaluate standard setting for specialist physician examinations in South Africa. This was done by:

- Determining the *knowledge, attitudes, views and perspectives* of CoP examiners with regard to standard setting, in order to provide appropriate training for examiners of the CoP.
- Determining and comparing the *performance* (pass marks and failure rates) of the Cohen method, to that of the more established Angoff method, in setting pass standards for postgraduate written specialist examinations in the CoP.
- Determining and comparing the *utility* of the Cohen method, to that of the more established Angoff method, as a standard setting method for routine use in postgraduate written examinations of the CoP.

Subsequently, conclusions and related recommendations will be made to achieve the goal of the study.

1.4.3 Objectives of the study

The objectives of this study are listed below together with a brief description of how each objective was addressed. A more detailed description of the methodology used to address each objective is provided in Chapter 3.

1.4.3.1 *Conceptualise the role of standard setting as it pertains to assessment in medical education and contextualise it to postgraduate written assessments for specialist physicians in South Africa.*

This was done by conducting a thorough review of the medical education literature on the topic of standard setting as well as a document analysis of assessment and standard setting regulations/policies relevant to this study. This objective addressed Research Question 1.

1.4.3.2 *Determine the knowledge, attitudes, views and perspectives of the CoP examiners about standard setting*

This was done by conducting a situational analysis at the start and end of the study using an online questionnaire survey. The survey was based on key issues identified in the literature review and informal discussions with examiners of the CoP. This objective addressed Research Question 2 and also made a contribution towards addressing Research Question 3.

1.4.3.3 *Design, deliver and evaluate the impact of a seminar dealing with standard setting in the CoP*

The content of this seminar was based on the results of the initial situational analysis conducted at the start of the study. This objective addressed Research Question 2 and also made a contribution towards addressing Research Question 3.

1.4.3.4 *Determine the performance (pass marks and failure rates) of the Angoff method of standard setting using five cycles of written FCP (SA) examinations data*

This was done by using the original Angoff method, as explained by Downing *et al*. (2006:53), to determine the pass marks and failure rates for five cycles of written FCP (SA) examinations data. This objective addressed Research Question 3.

1.4.3.5 *Determine the performance (pass marks and failure rates) of the Cohen method of standard setting using the same five cycles of written FCP (SA) examinations data as in 1.4.3.4*

This was done by applying and modelling the Cohen method (Cohen-Schotanus & Van der Vleuten 2010:157) to determine the pass marks and failure rates for the *same* five sets of the written FCP (SA) examinations data as described in 1.4.3.4. This objective addressed Research Question 3.

1.4.3.6 *Determine the variability of the scores of the top-performing candidates, in order to substantiate the assumption within the Cohen method that the test scores of the top candidates sitting the FCP (SA) written examinations, are stable and, therefore, comparable*

This was done by analysing the scores of the top performing candidates to determine the stability of their performance over five cycles of three different formats of written FCP (SA) examinations. This objective addressed Research Question 3.

1.4.3.7 *Use the findings of Objectives 1.4.3.2 – 1.4.3.6 to contribute to the evaluation of the utility (as defined in Chapter 2) of the Cohen method, as compared to the Angoff method, for the written FCP (SA) examinations*

This was done by incorporating the findings from objectives 1.4.3.2 – 1.4.3.3 (the change management process) with the statistical analysis on the results of objectives 1.4.3.4 – 1.4.3.6 (a comparison of the performance of the two methods), to synthesize and determine the results of this objective. This objective addressed Research Question 3.

1.5 DEMARCATION OF THE FIELD AND SCOPE OF THE STUDY

This study is positioned in the field of health sciences education, specifically addressing the process of setting pass standards for postgraduate written certification examinations of specialist physician trainees in South Africa. For reasons previously

articulated (section 1.2.3), it can be further described as dealing with quality assurance of the written assessment components of the Fellowship of the College of Physicians of South Africa examinations or the FCP (SA) examinations.

This study set out to determine the knowledge, views, attitudes and perspectives on standard setting of the examiners of the CoP and how this changed in response to training about standard setting and 'hands-on' experience in setting standards for the written components of the FCP (SA) conducted between 2011-2014. Secondly, the study also evaluated and compared the outcomes of both the Angoff and Cohen methods of standard setting using five cycles of examination data for each of the three written components of FCP (SA) examinations, conducted in the period August 2011 to March 2014. The examination cycles are explained in more detail in Chapter 3.

In a personal context, the researcher in this study qualified as a medical doctor in 2000 (MB.,Ch.B) from Stellenbosch University, South Africa and went on to obtain a Master's degree in Medical Education (M.MEd.) at Dundee University, Scotland. He has gained working experience in medical education in the United Kingdom and South Africa. He is currently employed as a clinical educationalist in the Department of Internal Medicine at the University of the Free State (UFS), South Africa. He is also a part-time tutor on assessment and standard setting at the Centre for Medical Education, Dundee University, Scotland.

1.6 THE VALUE AND SIGNIFICANCE OF THE STUDY

This study can be seen as a response to the call for further research on standard setting and evaluation of high-stakes postgraduate licensing examinations made by Burch and colleagues after studying the Part II FCP (SA) examinations (Burch & Norman 2009:446; Burch *et al.* 2008:532). These papers are discussed further in Chapter 2, section 2.3.5. In addition, Cohen-Schotanus and Van der Vleuten (2010:159) also called for more research on the Cohen method in other contexts.

Furthermore, it was of great value to determine the extent to which standard setting was accepted and valued by the examiners of the CoP and what training was needed to enable the examiners to endorse and adopt a credible and defensible way of setting pass standards for CMSA examinations. The study provided the CoP with empirical data demonstrating the impact of existing regulations on pass marks and failure rates

and the effect of different methods of standard setting on the same data. This study will enable and empower the CoP, and the wider CMSA, to review policy and regulatory decisions about pass standards and the use of formal standard setting processes.

If the Cohen method proves to be explicable, defensible, stable, and acceptable it will improve the quality of postgraduate assessment within the national postgraduate examining body. It will potentially also aid and encourage further quality assurance research with regard to standard setting in other colleges in the CMSA and perhaps also in undergraduate medical programmes in South Africa.

A recent study conducted in South Africa noted the "lack of awareness of the results and process of standard setting in the country" (Pitoniak & Yeld 2013:23). Although this was a concerning statement, it was not surprising. Formal standard setting processes are not currently part of all medical education assessment strategies.

Therefore, in summary, probably the greatest value and contribution of this research, and the subsequent publications from this Ph.D. study, are the extent to which it will contribute to raising awareness and stimulating change towards improving the quality of the pass/fail decision making in medical education assessment in South Africa. This study may facilitate the setting of explicable, defensible, stable and acceptable pass standards in all the medical schools in South Africa and the CMSA, as well as possibly introduce the Cohen method as a sustainable option towards implementing standard setting. This will be a positive and progressive quality assurance step for medical education and healthcare provision in South Africa.

This potential improvement in defensibility of passing standards is of particular significance since the literature (Barman 2008:961) warns against the potential legal ramifications and indefensible positions of medical education institutions or examining bodies not employing robust and rigorous standard setting processes, with particular mention of the unacceptable traditional practice of the fixed percentage pass mark. The importance of a 'documented process' as an integral part of rendering the derived pass mark as defensible, from a legal perspective, was also echoed by other authors (Carson 2001:428).

1.7 RESEARCH DESIGN OF THE STUDY AND METHODS OF INVESTIGATION

To ensure reliability and validity of a research project one has to make sure that the methods followed are clear to the readers. Therefore, the design and methods are explained in detail in Chapter 3 of this thesis. A brief synopsis is given here, to orientate the reader.

1.7.1 Design of the study

The concept of a research design is synonymous with the 'blueprint' of a study, describing the research and data collection processes. This study was about improving practice by introducing change in an educational context, while also simultaneously studying the results, impact and effects during the 30 month change management process to date. As a result, an *action research* framework and approach (Riel 2010:online) was used for the two research components of this study. An outline of the research process is provided next and is schematically presented in Figure 1.3.

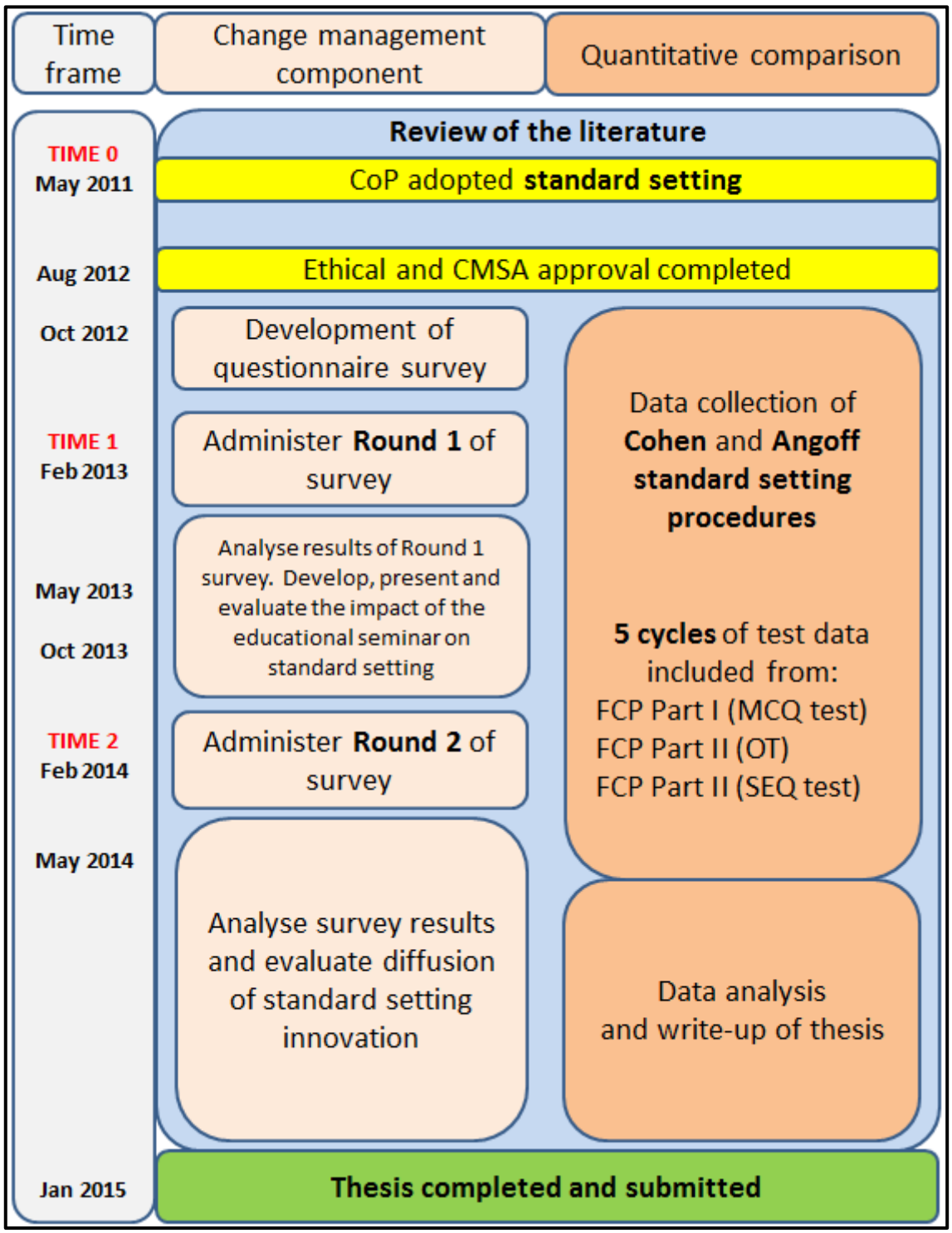


FIGURE 1.3: A SCHEMATIC OVERVIEW OF THE STUDY
[Compiled by the researcher, SCHOEMAN 2014]

1.7.1.1 *Literature review*

A thorough review of the literature was undertaken to form a comprehensive understanding of the concept and context of standard setting in health professions education, and specifically in medical education. In addition, a document analysis of

local policies and regulatory systems was also done to further inform the context in which this study took place. These actions addressed Research Question 1 (as well as objective 1.4.3.1).

The key themes emerging from the literature review having a bearing on the research goal, aims and objectives include (from Figure 2.1):

Conceptualisation:

- The role of assessment in medical education;
- The importance of high quality assessment processes to enable standard setting;
- Concept, principles, classification and methods of standard setting;
- The social accountability of standard setting;
- The utility of standard setting methods;

Contextualisation:

- International perspective on standard setting in medical education;
- South African higher education perspective;
- South African medical education perspective;
- Postgraduate medical specialist education in South Africa;
- Standard setting for specialist physicians in South Africa.
- Change management and the diffusion of innovation

As mentioned in section 1.1.1, this study had two research components, each with its own methodological design. Full details of the methodology of the individual components are provided in Chapter 3. A brief summary of each component of the study is provided here.

1.7.1.2 *The first component (the prospective cohort study)*

Research Question 2 and objectives 1.4.3.2 to 1.4.3.3 relate to the knowledge, attitudes, views and perspectives about standard setting of the 2010 – 2012 examiners of the CoP. For this component of the study, a prospective cohort study design was employed and the examiners (n=54) were asked to complete an online questionnaire (Appendix A-1, attached at the end of the thesis, with all the other Appendices A-F) at the start of the study (Time 1 – February 2013), which took place 18 months after standard setting had been implemented.

The questionnaire contained 16 questions or statements which the participants were asked to respond to or rate on a 5-point Likert scale, after the questionnaire had been piloted and adapted. After each question or statement the participant had the opportunity to add free text comments. The data from round one of the questionnaire were analysed to determine the position of the CoP examiners, regarding standard setting, and to identify their learning needs in order to design a standard setting seminar (the intervention).

Round 2 of the questionnaire survey (Appendix A-2), administered towards the end of the study (Time 2 – February 2014), was used to measure the impact of the seminar and further exposure to the standard setting process (in combination), on the same examiners' knowledge, attitudes, views and perspectives about standard setting. Their views on the long-term feasibility and sustainability of the Angoff method were also evaluated in this round.

1.7.1.3 *The second component (comparative study)*

For this component of the study, the utility of the Cohen and Angoff methods of standard setting, for the written components of the FCP (SA) examination, were determined and compared. The findings address Research Question 3 and objectives 1.4.3.4 to 1.4.3.7.

Five cycles of test data, from the three different written components of FCP (SA) Part I and II examination, were used to compare the two methods. The performance of top candidates in each cycle was also compared to test the hypothesis, upon which the Cohen method is based, that top candidates in successive examination cohorts perform in a stable manner over time, if test difficulty is similar.

The detailed description of the population, sampling methods, data collection techniques, data analysis and reporting, and ethical considerations are provided in Chapter 3.

1.8 IMPLEMENTATION OF THE FINDINGS

The findings and recommendations of this study will be submitted to the CoP council as well as to the CMSA Senate, for consideration and discussion. The findings will also be submitted for publication to international medical education journals.

1.9 ARRANGEMENT OF THE THESIS

The research report and the final outcome of the study are arranged as follows:

Chapter 1: Orientation to the study.
In this chapter, the background to the study as well as the concept and principles of standard setting were outlined. In addition, the problem statements, research questions and study hypotheses were stated. The overall goal, aim and objectives of the study, as well as the demarcation of the field and scope of the study, the significance and value of the study were also provided. In conclusion, the research design and methods of the study, the envisaged implementation of the findings as well as an overview of the lay-out of this Ph.D. report were outlined.

Chapter 2: Standard setting and assessment in medical education.
This chapter provides a review of the literature as it pertains to the conceptualisation and contextualisation of assessment and, in particular, standard setting in medical education. The published literature on standard setting and a document analysis of relevant regulations and policies, as pertaining to this study's scope, are discussed.

Chapter 3: Research design and study methodology.
The research design and methodology that was used in each of the research components of this study is described in more detail in this chapter.

Chapter 4: Introducing standard setting in the College of Physicians of South Africa – a process of change and diffusion of innovation.
In this chapter, data derived from evaluating the knowledge, attitudes, views and perspectives of the CoP examiners, over a 30 month change management process, relating to the introduction and gradual implementation of standard setting in the CoP, are presented and discussed. Three data gathering events took place to evaluate the

knowledge, attitudes, views and perspectives of the CoP examiners towards standard setting, over this time period:

1. The initial situational analysis (Time1 – Feb 2013) – measured the impact of 18 months of exposure to the Angoff and Cohen standard setting methods;
2. A training and educational seminar based on the results of the initial survey;
3. The final situational analysis (Time2 – Feb 2014) – measured the impact of the seminar and 30 months of exposure to the Angoff and Cohen standard setting methods.

The chapter provides a picture of the change in knowledge, attitudes, views and perspectives of the CoP examiners towards standard setting, as well as their views on the utility of the Angoff and Cohen methods.

Chapter 5: FCP (SA) Part I Multiple Choice Question (MCQ) Test - Comparing the performance of the Angoff and Cohen methods

In this chapter, a comparison of the Angoff and Cohen standard setting results (pass marks and failure rates), as it pertains to five cycles of data collected from the FCP (SA) Part I MCQ test, is presented and discussed.

Chapter 6: FCP (SA) Part II Objective Test (OT) - Comparing the performance of the Angoff and Cohen methods

In this chapter, a comparison of the Angoff and Cohen standard setting results (pass marks and failure rates), as it pertains to five cycles of data collected from the FCP (SA) Part II OT, is presented and discussed.

Chapter 7: FCP (SA) Part II Short Essay Question (SEQ) Test - Comparing the performance of the Angoff and Cohen methods

In this chapter, a comparison of the Angoff and Cohen standard setting results (pass marks and failure rates), as it pertains to five cycles of data collected from the FCP (SA) Part II SEQ test, is presented and discussed.

Chapter 8: Overall discussion and conclusions of standard setting for specialist physician examinations in South Africa.

In this chapter, overall deliberation of the findings and implications of the entire study, drawn from the discussions in the preceding four chapters, are made and discussed.

In addition, conclusions based on the findings and discussion of the two different research components, are drawn. Practical and future research *recommendations*, as well as the *limitations* of the study, will also be described and discussed.

1.10 CONCLUSION

This chapter has provided the background and introduction to this report on a Ph.D. study investigating the introduction, implementation and evaluation of standard setting for specialist physician examinations in South Africa. The study set out to contribute towards improving the quality (explicability, defensibility, stability, and acceptability) of the FCP (SA) written examination processes as a crucial component of quality assurance in the training of specialist physicians in South Africa. This study and its outcomes may serve as a guide to other organisations wishing to implement standard setting processes for undergraduate or postgraduate medical training programmes in South Africa.

CHAPTER 2

STANDARD SETTING AND ASSESSMENT IN MEDICAL EDUCATION

2.1 INTRODUCTION

The aim of this chapter is twofold. Firstly, to review and discuss the literature relevant to the *conceptualisation* of standard setting as it pertains to assessment in medical education. The second aim of the chapter is to review and discuss the literature relevant to the *contextualisation* of the study within the domain of standard setting as it relates to assessment in the context of the postgraduate specialist certification examinations offered by the Colleges of Medicine of South Africa, and specifically the CoP.

The chapter opens with a discussion of the meaning of the term 'standard setting' and its role in assessment in medical education. A focussed discussion about the purposes of assessment in medical education, the importance of, and key elements of high quality assessment processes, as well as different types of assessment relevant to this study, is also provided.

The discussion then moves on to review the concept, importance and underpinning principles of standard setting, together with a classification of different standard setting methods. Different standard setting methods commonly used are briefly reviewed, followed by a detailed review and discussion the Angoff and Cohen methods, respectively. The social accountability role of standard setting, as a quality assurance process in healthcare workforce production, is discussed, since this study and thesis relates directly to the production of the specialist physician workforce in South Africa. This part of the chapter ends by describing the utility parameters of standard setting methods, as derived from the literature. The parameters are organised and presented in a usable framework, which is used in Chapter 8 to evaluate the utility of the Cohen method, as compared to the Angoff method, in the context of the CoP, using the data reported and discussed in Chapters 4 – 8 of this thesis.

While the first part of the chapter explains and discusses the concept and role of standard setting in medical education assessment, the second part provides the context, rationale and motivation for this research project. It starts with an

international perspective about standard setting in medical education and gradually zooms in on the South African context and scope on this project, i.e. the written components of the licensing examination for specialist physician trainees in South Africa.

For the purpose of this thesis, a brief overview of the regulatory system of higher education in South Africa, of which medical education is part, is provided. This includes a description of the regulatory functions of the HPCSA and the CMSA, which was recently appointed by the HPCSA to administer the exit licensing examinations for all medical and dental specialist trainees who started their training on, or after, 1 January 2011 (HPCSA 2011:1). The CoP, one of the 28 constituent Colleges of the CMSA and the focus of this thesis, also has its own internal regulations about assessment and these were reviewed in terms of the process by which pass marks for their written assessments were determined, prior to the introduction of formal standard setting as described in this thesis. At the time of embarking on this project the CMSA did not have any overarching assessment regulations publicly available on its website to incorporate into this review (CMSA 2015:online).

Before the implementation of standard setting in the CoP, the knowledge, views, attitudes and perspectives of the CoP examiners about standard setting were not known. Hence, the need arose to also review published data describing the knowledge, views, attitudes and perspectives of examiners/educators/assessors about standard setting prior to, or soon after, the introduction of standard setting procedures in other contexts. Therefore, a short section of the chapter is devoted to a review of the wider role and purpose of licensing examinations, specifically literature describing standard setting processes for written assessments in this context.

Since the project described in this thesis involved a process of introducing major change to high-stakes written assessment practices, the final section of the chapter is devoted to reviewing and discussing the relevant literature regarding change management and the 'diffusion of innovation' through a social system, specifically in the context of this project and medical education.

Figure 2.1 provides a useful summary of the key elements of the literature review and document analysis reported in this chapter. The work described in this thesis is situated at the interface between the *conceptualisation* of standard setting, as

articulated in the literature, and the *contextualisation* of standard setting, i.e. postgraduate specialist certification examinations, which are mandated by the HPCSA and conducted by the constituent Colleges of the CMSA. As shown in Figure 2.1, change management and the diffusion of innovation were essential features of the study described in this thesis.

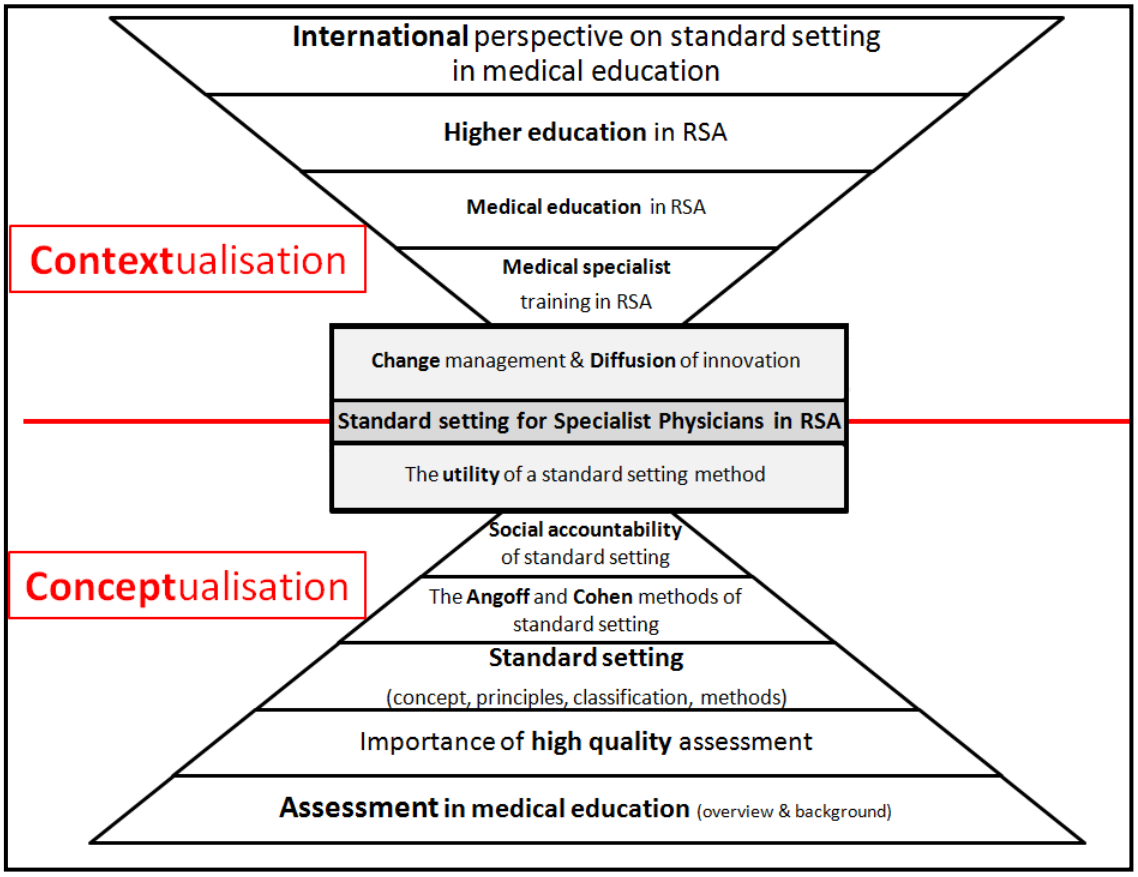


FIGURE 2.1: CONCEPTUALISATION AND CONTEXTUALISASION OF STANDARD SETTING FOR SPECIALIST PHYSICIANS IN RSA
[Compiled by the researcher, SCHOEMAN 2014]

2.2 CONCEPTUALISATION OF STANDARD SETTING IN MEDICAL EDUCATION ASSESSMENT

2.2.1 The purpose of medical education and assessment

To fully understand the concept of standard setting, in particular as it relates to assessment of student learning in medical education, the relationship between medical education, assessment and standard setting has to be explained. Medical education, in its broadest context, aims to enrich individuals through curricula comprising of knowledge (cognitive abilities), skills (psychomotor abilities) and attitudes (affective

abilities) using a wide variety of teaching and learning strategies and processes (Burch 2007:3; Epstein, Cox & Irby 2007:387; Shumway & Harden 2003:571; Swanwick & Buckley 2010:xv; Tormey 2014:7). The purpose of this 'enrichment' is to equip medical trainees with the necessary abilities to enable them to reach professional competence and to practise effectively and safely as a medical practitioner (Epstein *et al.* 2007:388). In theory, this process is never completed, but at some stage of this learning and training continuum, the need arises from society and the individual to assess the knowledge, skills and attitudes that have been acquired to enable certification as a qualified professional who has been judged as safe and competent to practise and serve the public (Epstein *et al.* 2007:388). There are, of course, different levels in this professional development process in terms of the level of knowledge and skills acquired, as illustrated in Figure 2.2.

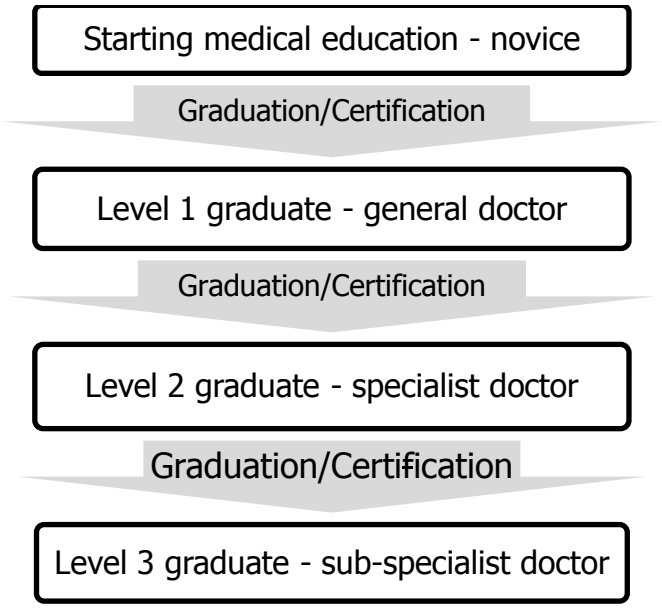


Figure 2.2: CAREER PROGRESSION IN THE MEDICAL PROFESSION [Compiled by the researcher, SCHOEMAN 2014]

From the preceding discussion, it is clear that there has to be an overarching organisation that governs (regulates) the certification of medical practitioners at all the levels shown in Figure 2.2, which assures the public of the *quality* of this education and training process. In this context, the word 'quality' means that the medical practitioner has met the standards of knowledge, skills and attitudes at the required level, and is deemed a safe and competent practitioner at the level of certification awarded. This quality assurance process is at the core of building a reputable and trusted body of professionals anywhere in the world (Epstein *et al.* 2007:394).

As explained, quality assurance at all levels of certification is of paramount importance. For this reason, assessing the achievement of the learning outcomes of training programmes in medicine and determining the relevant levels of mastery of the necessary knowledge, skills and attitudes of medical trainees, form an integral part of the quality assurance process (Epstein *et al.* 2007:388; Roberts, Newble, Jolly, Reed & Hampton 2006:535; Schuwirth & Van der Vleuten 2010:195; Shumway & Harden 2003:578).

All accredited medical training institutions and national examining bodies for postgraduate specialist certification are, in addition, charged by their respective regulatory authorities with the important task and responsibility of making decisions about the academic progression or graduation of their trainees or candidates. The quality, role and purpose of assessment in medical education are, therefore, of key importance, and have been well described in the literature. The three key purposes of assessment can be summarised as follows:

1. The measurement of student learning or mastery in the subject to make decisions about progression or certification (summative assessment);
2. The use of assessment results to provide feedback to learners about their mastery of the subject in order to facilitate and stimulate further learning (formative assessment);
3. The use of assessment data to drive curriculum change to improve the training of learners (Burch 2007:1-2; Jolly 2010:209; Southgate, Hays, Norcini, Mulholland, Ayers, Woolliscroft, Cusimano, McAvoy, Ainsworth, Haist & Campbell 2001:475; Wass, Van der Vleuten, Shatzer & Jones 2001:945).

The non-summative roles of assessment relate to the educational impact of assessment on the individual learner and the educators, who are in charge of the curriculum, respectively. These two aspects are beyond the remit of this thesis and are not discussed further. The use of assessment to make decisions about progression or certification is relevant to standard setting and is discussed next.

The notion of what constitutes "good enough to pass" (Bandaranayake 2008:836; Cusimano 1996:S112) is the basic question medical educators have to answer in order to make a decision about the progression of their students or trainees. To make such a decision, there are a number of important factors which influence the quality of the

assessment data and need to be acknowledged. The definition of quality in this context refers to the trustworthiness and accuracy (usefulness) of the data as a representation of the trainee's learning or mastery of the assessed domain of knowledge, skills and/or attitudes.

Van der Vleuten (1996:54) describes various factors which influence the utility (quality) of assessment instruments, which aim to measure competence in medical education, and their results. They include choosing the appropriate assessment instrument for the assessed construct bearing in mind its related reliability and validity credentials, as well as other factors relating to the cost-effectiveness, educational impact and acceptability of the chosen instrument among the relevant stakeholders.

A more detailed discussion on the importance of high quality of assessment and results for standard setting and what constitutes as quality in assessment, follows later in this chapter.

2.2.2 Assessment strategies

In 1990, Miller (1990:S63) described his hierarchy of assessment in terms of four levels of understanding and performance, as a learner progresses from a novice to an expert. The well-known four-layer pyramid is illustrated in Figure 2.3.

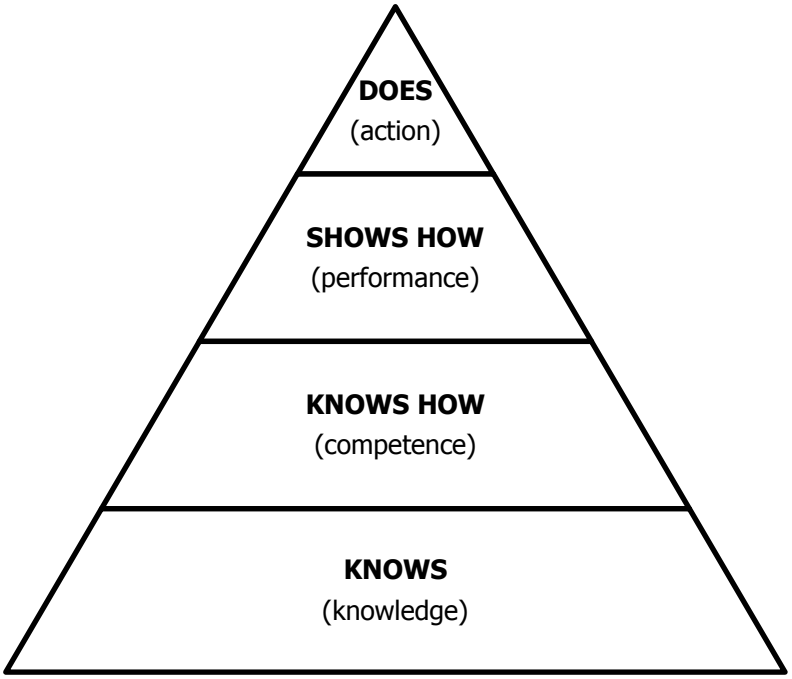


FIGURE 2.3: MILLER'S PYRAMID OF ASSESSMENT HIERARCHY
[Figure taken from Miller (1990:S63)]

Based on Miller's pyramid, summative assessment strategies in medical education can be divided into two broad assessment strategy groups: 1) Written or oral assessment of theory (knowledge) and 2) Performance assessment of psychomotor skills and attitudes (Wass *et al.* 2001:946).

The pyramid's simple description of the development of competence expressed as a function of the purpose of a particular level of assessment, has become a popular and practical classification of assessment instruments in the medical education literature. For example, true/false MCQs, which typically assess recall of facts, operate on the "knows" level (Case & Swanson 1998:18; Wass *et al.* 2001:947), which has the broadest base.

The next level, "knows how", is where application and processing of knowledge can be tested using clinical case-based, single best-answer MCQs (Case & Swanson 1998:18; Van der Vleuten & Schuwirth 2005:313; Wass *et al.* 2001:947).

The third level, "shows" requires that students demonstrate their knowledge and skills to using tasks, for example an Objectively Structured Clinical Examination (OSCE) (Schuwirth & Van der Vleuten 2010:200; Wass *et al.* 2001:947).

The top level, "does", which is equivalent to the "real world" context of practice, measures the student's ability to perform tasks in the authentic workplace setting (Norcini & McKinley 2007:245-248; Schuwirth & Van der Vleuten 2010:201; Wass *et al.* 2001:948). Burch (2007:6) describes these bottom three levels as "in-vitro" assessment and the top level as "in-vivo" assessment, which is also a relevant and appropriate description of the development of the context within which higher order assessments occur, from the classroom to the workplace (Southgate *et al.* 2001:475).

The work of this thesis is restricted to the written assessments of knowledge and the application of knowledge as used in the FCP (SA) examinations of the CoP and how these assessment results are used to make pass/fail decisions about the mastery of the discipline of Internal Medicine upon entry (Part I) and at the time of completion, i.e. certification (Part II). Therefore, the assessment of cognitive skills (knowledge) using written assessment tools will be discussed in the next section. The assessment of clinical competence at the levels of "shows how" and "does" is outside the scope of this thesis and not discussed any further.

2.2.3 Written assessment of theoretical knowledge

The literature describes a range of methods to assess the leaning of theoretical knowledge (Jolly 2010:211-213). This section focuses on the assessment methods relevant to the work reported in this thesis. These knowledge assessments are located in the lower levels of Miller's pyramid, i.e. the "knows" and "knows how" levels. There are essentially two assessment instrument formats used to assess knowledge. They are selected-response formats and constructed-response formats. Each type is discussed here in terms of their basic characteristics and later in the chapter the psychometric properties of these instruments are discussed (section 2.2.4.3).

Selected-response formats

This format includes two types of question items, namely multiple choice questions (MCQs) and the more recently developed, extended matching items (EMIs) (Epstein *et al.* 2007:390; Jolly 2010:212). Both are renowned for their psychometric rigour and efficiency to test large volumes of knowledge (Downing 2009b:L2382; Norcini *et al.* 2011:208). Previous authors have highlighted the fact that it is the content and format of the stimulus provided in the MCQ that determines the level on which the question operates in terms of Miller's pyramid (Wass *et al.* 2001:947).

Clinical case-based, single best-answer MCQs with 3-5 possible options, are more suitable than traditional true/false MCQs to test higher order cognitive skills, such as interpretation, analysis, application and synthesis of knowledge – i.e. Bloom's taxonomy level 3 and above (Case & Swanson 1998:18). In fact, the well-respected National Board of Medical Examiners (NBME) in the United States of America (USA), who administer their national licensing examinations (USMLE step 1 and 2) for basic medical education (MD), has stated in their most recent manual on constructing high quality MCQ items, referring to true/false MCQs, that: "We find that, to avoid ambiguity, we are pushed toward assessing recall of an isolated fact — something we are actively trying to avoid. We find that application of knowledge, integration, synthesis, and judgement questions can better be assessed by one-best-answer questions. As a result, the NBME has completely stopped using true/false formats in its examinations" (Case & Swanson 1998:18).

This is in keeping with the recent trend in medical education to exploit the powerful influence of assessment on learning, in order to promote learning styles that facilitate the development of clinical reasoning and problem-solving skills, which are required of graduates in the clinical context of patient care and management (Schuwirth & van der Vleuten 2011:793). This approach is referred to as 'assessment for learning' as opposed to the traditional 'assessment of learning'. This approach recognises and incorporates the effect of assessment on the learning approaches and behaviour of students and trainees (Schuwirth & van der Vleuten 2011:793; Tormey 2014:3).

As mentioned previously, it is the content and format of a question's stimulus that is the critical factor, not the response format, which determines the level of reasoning required of a candidate (Schuwirth & Van der Vleuten 2010:199; 2011:783). This implies that the format in which the response of a candidate is captured, either by selecting from a list or by self-generating an answer, has less influence on the thinking and learning of the candidate, when compared to the question stimulus (Schuwirth & Van der Vleuten 2010:199).

EMIs have in recent years become an attractive extension of case-based single best-answer (SBA) MCQs with 3-5 possible options. They are based on the same construction principles as case-based SBAs, but differ in two important ways. The cases used in EMIs are all on the same theme, for example, heart valve lesions, and secondly, the list of options, applicable for all the cases included in the EMI, typically has 8-12 options, depending on the theme of the EMI and the constrains of the data capturing device used. This makes the possibility of guessing the correct answer, which is one of the largest criticisms against MCQs, far less of a concern for test administrators.

The good 'utility' credentials (reliability, validity, educational impact, acceptability and cost-effectiveness) of SBA MCQs and EMIs have made them increasingly popular and propagated written assessment instruments (Van der Vleuten & Schuwirth 2005:309-311) to assess not just knowledge recall (knows), but also analysis, interpretation and application of knowledge in the clinical context ('knows how') (Wass *et al.* 2001:947).

Constructed-response formats

This method of testing has been used in medical education for centuries (Downing 2009b:L2257). Classic essay-type questions were the mainstay of medical education written assessments for many decades (Norcini *et al.* 2011:208) and continues to be used in organisations and institutions due to their perceived face validity (Wass *et al.* 2001:947). They are regarded as testing factual recall, application of knowledge and clinical reasoning in an in-depth manner, as well as providing insight into the candidate's ability to construct a sound argument when answering the question. However, there were some serious concerns about this instrument which motivated the majority of educators, institutions and organisations to phase it out of their assessment systems (Wass *et al.* 2001:947). The challenges included:

- Lack of psychometric rigour, mostly due to a lack of sufficient sampling from the assessed domain (Downing & Haladyna 2004:328; Norcini *et al.* 2011:208);
- Concerns about ambiguity of questions with a lack of sufficient focus (multiple ways to interpret and approach the question) in the way questions were phrased in order to produce a single, agreed-upon model answer (Downing 2009b:L2285; Jolly 2010:210-211);
- Testing inefficiency and resource implications when assessing large groups of candidates (Jolly 2010:210-211).

These limitations led to the development of shorter essay-type questions, referred to as short answer questions (SAQs). SAQs require shorter written responses and although many improvements were made in terms of psychometrics (improved sampling), model answers and test efficiency remain concerns. The content of the SAQs, as opposed to the format, remains the predominant factor determining what cognitive level of Bloom's taxonomy is being addressed (Jolly 2010:216).

2.2.4 The importance of high quality assessment data

To make good and fair decisions and inferences about students' abilities, high quality assessment data are needed (Crocker & Zieky 1995:19; McGaghie, Butter & Kaye 2009:L2914). This section discusses what constitutes high quality assessment data, since this has significant implications for standard setting in medical education. Decisions about which candidates pass or fail a high-stakes written assessment are based on performance data. Therefore, if the performance data from a test are of

poor quality (invalid or unreliable for the high-stakes purpose of the assessment) then poor quality progression decisions are likely (Downing 2004:1007; McGaghie *et al.* 2009:L2921). Good decisions cannot stem from bad assessment data (Crocker & Zieky 1995:19). High quality standard setting starts with and relies on high quality performance data, generated by high quality assessment methods and strategies which are appropriately selected and developed for the purpose of the assessment (McGaghie *et al.* 2009:L2921).

It is important and useful to define and clarify the meaning of some commonly used terms in this section and in the thesis. *Measurement* is the process of assigning a numerical value to quantify the construct being measured (Tavakol & Dennick 2011b:448). *Assessment* is the measurement of learning (Tavakol & Dennick 2011b:448), which is carried out by asking candidates to sit tests (or assessments). Tavakol & Dennick (2011b:449) explain that tests are considered *objective* if they are administered, scored and interpreted independently, without subjective judgement of raters. The qualitative grading of constructed-response assessments is briefly discussed further in section 2.2.7.1.

For the performance data of educational measurements to be regarded as high quality (trustworthy and credible), they need to be *valid and reliable* (McGaghie *et al.* 2009:L2914). These two important concepts, validity and reliability, are reviewed and discussed in the remaining part of this section.

2.2.4.1 *Validity of test data*

This is the single most important topic in assessment according to Downing and Haladyna (2009:L466). Questions such as: "How valid is the inference we make from the results of a test?" and "Do the scores represent what we want to know?" reflect the essence of what validity is all about. The validity of a test result is, therefore, defined as the evidence that supports or refutes the *meaning* assigned to the test result (Downing 2003b:830).

The contemporary view from the literature is that validity is regarded as a *unitary* concept taking into account multiple sources of evidence, including reliability (Downing 2003b:830). Downing (2003b:831) goes on to explain that all validity is now regarded as aspects or components of *construct* validity. A construct is an abstract concept or

trait which a test is attempting to measure and what the results aim to numerically represent (Downing 2003b:831). The evidence provided to support this claim either validates or erodes the meaning associated with the test results (Downing 2003b:830).

The *purpose* of the assessment is therefore of primary importance, since it defines the construct that the test aims to measure (Schuwirth & Van der Vleuten 2010:196). Validity evidence, therefore, needs to *link or support* the purpose of the test with the meaning of the results (Schuwirth & Van der Vleuten 2010:196).

In addition to the purpose of the test, is the *quality* of a test, which is inherently linked to validity (Rodriguez 2005:6). A high quality test will provide evidence in support for the validity of the test's results. Tests consist of items and hence the *number of items* included in the test together with the *quality of the items* are significant sources of validity evidence for a test's results (Downing & Haladyna 2009:L549).

The greatest threats to the validity of test results can be summarised in two themes - Construct under-representation (CU) and Construct-irrelevant variance (CIV). The CU of a test relates to insufficient and/or inappropriate *sampling* from the test domain and CIV essentially stem from poor *item quality* factors (Downing & Haladyna 2004:38). Both of these validity threats influence a major component of validity – *reliability*. These key components of validity evidence each merit their own review and discussion, and is presented in the subsequent text.

2.2.4.2 *Number of items included in the test*

The number of items included in the test, the sample size, is a reflection of the representativeness of the test to the domain being tested (Tavakol & Dennick 2011b:450). For a test sample of items to be representative of the population of all the possible items covering the testing domain, the items in the test should represent the size (quantity) and the width of the domain (Hays, Hamlin & Crane 2014:2; Schuwirth & Van der Vleuten 2010:196). If there is not adequate (amount) and appropriate (width) domain representation in the sampling on the test, it will suffer a considerable validity threat, referred to as Construct under-representation (CU) (Downing & Haladyna 2004:328) as mention previously.

The previous paragraph explains the importance of *blueprinting* (Schuwirth & Van der Vleuten 2010:196), which is the test item mapping and selection process that proves that the items come from all areas of the domain under assessment. How many items from the blueprint are included in the test depends on the purpose and importance of the test. The greater the number of items from a wide selection of topics contained in the overall blueprint, the more valid and reliable the test *can* be. The 'can be' is important to note, since *width and number included* sets the potential or scope of what the specific test can attain in terms of validity and reliability (Tavakol & Dennick 2011b:450). Although the *width and number* of items included are not the only factors, they are probably the most important factors influencing the validity of a test's outcome (Cook & Beckman 2006:166.e10). For low stakes formative classroom tests, where the results will not have major significance, small numbers of items might be appropriate. However, for high-stakes assessments, such as the context of this thesis, the sampling needs to be *extensive* and *wide* to represent the large domain of Internal Medicine appropriately and to achieve the required reliability coefficient of 0.85 or above.

Although reliability is an *integral* part of validation evidence of test results (Rodriguez 2005:6), it is discussed separately in section 2.2.4.4.

2.2.4.3 *Item Quality*

An assessment item is the smallest unit of testing in an assessment. Therefore, it is the 'building blocks' of assessments and also the smallest unit of quality assurance (Downing 2009a:L1615; Hays *et al.* 2014:2). The quality characteristics of test items are important focus areas to develop and improve high quality tests (Downing & Haladyna 2009:L620; Rodriguez 2005:11; Tavakol & Dennick 2011b:451) and subsequently, produce high quality test *results* (data), which in turn is essential for a defensible standard setting strategy (Downing & Haladyna 2004:329).

Item format

As discussed previously in this chapter, items of written tests are broadly classified in two format types, namely selected-response (SR) items or constructed-response (CR) items. The *purpose* of the assessment primarily determines which format is most appropriate to use (Downing 2009b:L2245). As a general rule, it's not the test format that determines what construct can be assessed, but rather the test *content*

(Schuwirth & Van der Vleuten 2010:199; 2011:783; Van der Vleuten 1996:51). Choices of format should be guided by evidence and testing efficiency to ensure valid test outcomes (Downing 2009b:L2245).

Constructed-response (CR) test items are the only way educators can assess the *writing ability* of candidates (Downing 2009b:L2258). If this is a specific aim and purpose of the assessment, then constructed-response tests are appropriate. However, the amount of effort required to administer CR tests to ensure validity and reliability, greatly exceeds what is needed for SR tests. Furthermore, CIV is a significantly bigger problem in CR items, compared to SR test items, due to the added variable of *markers or judges* (Downing 2009b:L2265). The human element introduces subjectivity (bias) into the CR system, and effects such as tiredness, marking and counting errors, halo effects, personal views on content and stringency (hawk or dove effect) all compound to add significant CIV. For these reasons CR tests, especially longer essay-type tests that are laborious to mark, especially with large cohorts of candidates, have reduced validity and reliability (Downing 2009b:L2298) and therefore, support the move towards adopting SR tests (Ware & Vik 2009:238). Shorter CR test items allow more items per unit of test time (increased sampling), and reduce the CR tests' validity and reliability threats (Downing 2009b:L2291).

Selected-response (SR) tests and items, principally the multiple choice question (MCQ) formats are considered the true 'work horse' of the written assessment world and is backed up by over 90 years of research into its validity and versatility (Downing & Yudkowsky 2009:L245).

MCQs are widely regarded as the most useful written test format for assessing cognitive knowledge in medical education (Downing 2009b:L2369). This view is based on the wide construct testing ability of MCQs, their sampling and marking efficiency when testing large numbers of candidates and the high indices of validity and reliability they can generate.

Three options per MCQ item have emerged as the optimum number of options for most MCQ tests. Rodriquez (2005) made this recommendation after a meta-analysis of 80 years of research on the topic. He reported that reducing options from five to three had, on average, a reduction in item difficulty of 7% and no effect on item discrimination or reliability (Rodriguez 2005:10). Reducing options from four to three

had, on average, a reduction in item difficulty of 4%, increased item discrimination by 4% and increased reliability by 0.02 (Rodriguez 2005:10). Given the increased item writing efficiency and more items per test time possible (Rodriguez 2005:11), this is certainly a welcome evidence based finding.

In an age of increasing demand for transparency, the use of high quality testing systems to support or substantiate progression decisions about trainees, together with the need for cost-effectiveness for funding of educational organisations, contributes to the increased popularity of SR testing formats with testing agencies. From a conceptual perspective, these driving factors are well aligned with Van der Vleuten's (1996:54) description of the five *utility* parameters of an assessment (or testing) instrument.

The item quality factors discussed in the subsequent text are those which are determined by psychometric processes and collectively called *item analysis*. Item analysis data are a major source of the validity evidence of test results (Rodriguez 2005:6). They are primarily described from the SR (MCQ) perspective and comments are made about how this is translated to the CR context, where applicable.

Item difficulty

The item difficulty index reflects how hard (difficult) or easy the candidates found an item in a test (Sim & Rasiah 2006:68; Tavakol & Dennick 2011b:452). It could perhaps also be called the item easiness index, since the higher the value, the easier the item was for the cohort tested (Hingorjo & Jaleel 2012:143).

In SR formats, such as MCQ tests, item difficulty is reported as the *proportion of candidates* who selected the correct option for the item and is referred to as the Proportion Correct value (PC-value). The PC-value of an MCQ item is determined by summing the number of candidates who answered the item *correctly* and dividing it by the *total number* of candidates who sat the examination (Tavakol & Dennick 2011b:452). It is expressed as a value between 0-1. If 58% of the cohort marked the item correctly, the PC-value is 0.58. Both the PC-value and the converted PC-value percentage (%) are used throughout this thesis to refer to the difficulty of test items, based on the performance of candidates who sat the test. They have identical meanings, but are expressed on different scales.

The concept works slightly differently for CR items which are not scored in a binary (right or wrong) way. Item difficulty for CR items is calculated as the mean score of the candidates' performance on a particular item, and is called the item difficulty value (ID-value). If the cohort's mean score for an item was 45%, the item ID-value is 0.45. Both the ID-value and the converted ID-value percentage (%) are used throughout this thesis to refer to the difficulty of test items, based on the performance of candidates who sat the test. They have identical meanings, but are expressed on different scales.

Mean item PC-values of 0.72 were reported in the literature in a Norwegian study of undergraduate medical students' performance on four consecutive end of year MCQs papers (Ware & Vik 2009:240). Downing (2005:137) reported similar mean item PC-values of 0.71 in a study from the USA of undergraduate medical students' performance on four basic science MCQs papers. Items of moderate difficulty (PC-value/ID value 0.3 – 0.7) are psychometrically best, since they have a better chance of having good discrimination ability, because of the greater expected performance variance of the candidates on items that have moderate difficulty (Downing 2009a:L1691). Test items that are very easy or very difficult reduce the reliability of the test due to the fact that they cannot assess candidate variance (McManus, Mooney-Somers, Dacre & Vale 2003:611).

Item Discrimination Index (DI)

The DI of an item is its ability to discriminate between the top and the bottom performing candidates on the overall test. The classical method of determining the DI of an item is to calculate the difference between the mean performance on the item by the top and bottom 27% of the cohort in the overall test results (Sim & Rasiah 2006:69; Tavakol & Dennick 2011b:453). This method works well for large sample sizes (200+ candidates), since these large samples yield enough candidates in the respective top and bottom groups (Downing 2009a:L1702). The concern raised when this method has been used with smaller cohorts of less than 200 (Sim & Rasiah 2006:70), is the importance of having a large enough sample in the top and bottom groups to enable reliable DI calculations (Downing 2009a:L1702).

The use of 'thirds' (33% top and bottom groups) has been described in the literature as a compromise approach for smaller groups (Tavakol & Dennick 2011b:453). This approach was also used in the present study, since all the cohorts had fewer than 200

candidates. However, there are studies in the literature that used the 27% extreme group determination despite small cohort sizes; this casts some doubt over the reliability and validity of the reported DIs (Hingorjo & Jaleel 2012:144; Ware & Vik 2009:239).

The DI is expressed as a number between -1 and +1, where +1 means there is a 100% difference between the top group and the bottom group on the item, i.e. perfect discrimination. A value of -1 means 100% of the bottom group had the answer correct and 0% of the top group, which is the worst possible outcome for a DI (Tavakol & Dennick 2011b:453).

An acceptable DI for a test item is usually reported as value of 0.20 or greater (Downing 2009a:L4747), although some studies used values of 0.15 or more (Ware & Vik 2009:239) and others were more stringent, requiring values of 0.25 or greater (Hingorjo & Jaleel 2012:144). In this thesis, test items having a DI of 0.20 or greater were categorised as test items with good discrimination ability.

An example from the literature, which reported a mean DI for the test items of 0.19, was the same study by Downing (2005:137) described briefly in the item difficulty part of this section. Two studies also reported the mean percentage of test items which had good DIs. Ware and Vik (2009:240), from Norway, reported 65% and Hingorjo and Jaleel (2012:145), from Pakistan, reported 64%.

Test items that have poor DIs reduce the reliability of the test, because they are unable to assess candidate variance (McManus *et al.* 2003:611).

Other methods used to calculate item discrimination include the point-biserial (PBis) method (Sim & Rasiah 2006:70), a correlation-based method (Rodriguez 2005:6), which correlates the scores of the candidates on the item with their scores on the overall test (including the specific item) (Tavakol & Dennick 2011b:453). The literature contains a variety of these methods, but does not rank one superior to the other (Downing 2009a:L1670; Rodriguez 2005:6). They provide very similar estimates and are highly correlated, but use different methodologies (Downing 2009a:L1670; Rodriguez 2005:6).

Item distractor efficiency

The literature describes an effective MCQ item distractor (a wrong option) as one selected by 5% of the test cohort (Hingorjo & Jaleel 2012:143; Ware & Vik 2009:239). This is not a common item quality indicator since it is closely correlated to the DI of the item (Hingorjo & Jaleel 2012:145), but it may add additional useful information to improve distractors or remove non-functioning options and reduce the number of item options to three, regarded as the optimum number, based on evidence from a large meta-analysis (Rodriguez 2005:3-11). A Norwegian study by Ware & Vik (2009), which used best-of-five MCQs, showed on average, over four exam administrations, that only 17.5% of all their test items had three or four functional distractors. This supports the findings reported by Rodriquez (2005:11) .

Item writing flaws

Badly worded items, referred to as "flawed" items, add CIV to test scores, reduces test reliability and validity and make test items more difficult (Downing 2002:S104; 2005:141). Many high-quality manuals or guidelines for constructing MCQ items have been published in the literature and these were reviewed to produce a unified guide in 2002 (Haladyna, Downing & Rodriguez 2002).

The item writing manual form the National Board of Medical Examiners (NBME) in the USA (Case & Swanson 1998) is also highly regarded by the medical education community (Tavakol & Dennick 2011b:449) and overlaps extensively with the 2002 meta-analysis guidelines published by Haladyna and colleagues (2002:309-329). A detailed description of these criteria lies beyond the scope of this study, but failure to adhere to item writing guidelines has been shown to negatively affect the quality of test results and standard setting outcomes (Downing 2005:141).

Rater/marker effects

CIV challenges ascribed to raters/markers have been discussed earlier in the context of CR test items. Test-markers' subjectivity (hawk and dove or halo effect) and reader bias introduces significant error (CIV) into scores and present many validity threats to CR test results (Downing 2009b:L2341). In essay-type CR items *central tendency* marking, where most of the marks concentrate around the middle value of the items' weight, leading to reduced score variance and reliability, is a well-known validity and reliability concern (Downing 2009b:L2348).

Negative marking

Many authors have indicated that using forms of negative marking or formula scoring to discourage candidates form guessing in the test adds construct-irrelevant variance (CIV) or 'noise' to raw test scores (Downing 2003a:671). Although the concept of attempting to reduce the effect of guessing and achieve a more 'real' reflection of a candidates' true knowledge levels (i.e. removing noise) is understandable, the literature suggests that it does more harm than good from a psychometric perspective.

This is also a particularly important issue from a standard setting perspective. Two small studies by Downing (Downing 2002:S104; 2005:141), showed that the additional CIV introduced by negative marking adversely affected the pass/fail classifications of students. A more plausible alternative would be to stop correcting for guessing on the candidates' side (with negative marking or formula scoring of their test scores), and rather adjust the passing score upwards to account for the beneficial effects of educated guessing by lower performing and borderline candidates. This would not add CIV to the raw scores and hence would not negatively affect the validity or reliability of the test results. This strategy, described by Cohen-Schotanus and Van der Vleuten (Cohen-Schotanus & Van der Vleuten 2010:157) is used in the Netherlands. This approach is described in more detail in section 2.2.10 of this chapter, during the review and discussion of the Cohen method.

2.2.4.4 *Reliability of test data*

Reliability refers principally to the reproducibility of the test results from one sitting to the next (Gronlund 1998:210; McManus *et al.* 2003:609). It is an essential part of the validity evidence for a test, since test results that are not reliable, cannot be valid (Cook & Beckman 2006:166.e12; Downing 2003b:834; 2004:1007).

It originates from classical test theory (CTT) which states that:

Observed test score = True test score (T) + Construct-irrelevant variance (CIV) / error or 'noise' (Schuwirth & van der Vleuten 2011:788; Tavakol & Dennick 2011b:454)

The CIV could either be *systematic* (consistently have an influence such as too short time limit for a test) (Gronlund 1998:210; Tavakol & Dennick 2011b:454) or *random* (inconsistently has an influence such as rater bias or tiredness with marking test

answer-booklets) (Gronlund 1998:211; Tavakol & Dennick 2011b:454). Any CIV affects test scores and hence the validity of the meaning that is attached to the test results (Downing & Haladyna 2004:328-329). All tests have some CIV, but this must be kept to the absolute minimum by understanding which factors increase and decrease CIV in tests and addressing them in the test development process (Tavakol & Dennick 2011b:454).

Random CIV is the worst kind, due to its unpredictability. The effect of this unpredictable, random test 'noise' can, however, be measured and reported. This is what is referred to as the reliability measurement (Gronlund 1998:211). Reliability is typically reported using a *reliability coefficient*. Commonly used reliability measurements in written assessments include Cronbach's alpha (Rodriguez 2005:6; Tavakol & Dennick 2011a; b:455), Kuder-Richardson Formula 20 (Gronlund 1998:214; Rodriguez 2005:6; Ware & Vik 2009:239) and Generalisability Theory (Schuwirth & van der Vleuten 2011:789). The amount (%) of random CIV can then be calculated as:

$1 - (\text{the reliability coefficient})^2$ (Tavakol & Dennick 2011a:53)

The reliability coefficient helps educators judge how much variation they could expect if the test was taken under different conditions or at a different time.

For high-stakes tests, such as licensing assessments, high reliability coefficients are needed (to be a valid assessment) because of the significant implications of the results and hence, acceptable reliability coefficients are typically around 0.90 or above (Cook & Beckman 2006:166.e14; Downing 2004:1009; Tighe, McManus, Dewhurst, Chis & Mucklow 2010:2). The mean reliability coefficient reported in a 17-year review (1984 – 2001) of the MRCP (UK) Part I examination was 0.865 using the Kuder-Richardson formula 20 (McManus *et al.* 2003:610). A global review of published results of postgraduate licensing examinations between 1985 – 2000 reported reliability coefficients between 0.55 – 0.96, with a median of 0.77 (Hutchinson, Aitken & Hayes 2002:86). Hutchinson *et al.* (2002:74) suggested that 0.8 or 0.85 is the minimum reliability coefficient for high-stakes licensing testing. The literature does not have a consensus opinion on the accepted *minimum* reliability coefficient for high-stakes examinations, but a middle ground is in the order of 0.85.

Factors that reduce the reliability of test scores include: too few test items (Gronlund 1998:217), poor quality test items leading to a narrow range of test scores (Gronlund

1998:217), inadequate testing conditions (Gronlund 1998:217) and subjective marking, which leads to an increase in random errors (Gronlund 1998:217).

Several authors have warned about making inappropriate deductions about the reliability of a test purely on the basis of the reliability coefficient (Harvill 1991:181-182; Tighe *et al.* 2010:2). The reliability coefficient is influenced by the inherent measurement error (CIV) in the test as well as the variance in the scores of candidates. Therefore, if the same test is undertaken by a cohort of candidates with greater variance in ability (greater number of stronger and weaker candidates) the alpha coefficient will increase artificially, but the inherent test quality will remain unchanged (Harvill 1991:183; Tighe *et al.* 2010:2).

Tighe *et al.* (2010:3-5) demonstrated this effect in a large Monte Carlo simulation of 10 000 data points. They showed that administering the same test under the same conditions to the same 'candidates', who passed the test previously and were selected to sit it again, led to a 0.19 reduction in the reliability coefficient from 0.90 to 0.707 (Tighe *et al.* 2010:4-5). In the same paper, a similar effect was demonstrated in the 'real world' with a large review of the MRCP (UK) Part I and II written papers from 2002 - 2008. Only selected candidates who passed the Part I examination previously could attempt the Part II paper. The Part II cycles from 2002 to 2008 never reached the 0.90 reliability level even though it had 150 MCQ items and went through strict quality control measures. In 2004, the items were increased to 180 and in 2005 to about 270 to try and achieve a reliability of 0.90. This also failed. The mean reliability was 0.802. The reason for the inability to raise the reliability was due to the narrow performance band of the selected, rather homogeneous candidates sitting the Part II exam. The mean standard deviation (SD) over the time frame of the study was about 7%, which meant there was not enough variance in the performance to lift the α coefficient to 0.90 (Tighe *et al.* 2010:6-7).

Standard Error of Measurement (SEM)

However, the same authors showed that when controlling for the narrow SD, by using the SEM, which is a *direct* reflection of the test instrument itself and the measurement error contained within it, the Part II examination was in fact remarkably reliable, with a mean SEM of only 3.1% over the study period. They concluded that for high-stakes licensing examinations for selected, homogeneous cohorts of examinees with narrow ability ranges, the SEM is a better measure of test reliability than the reliability coefficient (Tighe *et al.* 2010:8). This view regarding the usefulness of the SEM as a

reliability indicator, was also suggested by Harvill in his 1991 paper explaining the SEM (Harvill 1991:181-189). This is particularly relevant in the context of this thesis, where the candidates attempting the Part II FCP (SA) written examinations are also relatively homogeneous, since they have all passed the entry-level Part I FCP (SA) MCQ test, were selected into residency programmes and have completed their residency training (or are nearing completion). Since the SEM is such an important topic in reliability, it merits further discussion.

The standard error of measurement is calculated from the reliability coefficient and the standard deviation of the test scores (Harvill 1991:182; Tavakol & Dennick 2011a:53; 2011b:456). It provides test administrators with an estimate of the range of standard measurement error around an *observed* score in the test (Tavakol & Dennick 2011b:456) and so it can be used to calculate the confidence intervals of students' *true scores*, given their observed scores on the test.

Harvill (1991:183) originally explained that the correct application of the SEM was to estimate the confidence intervals for where a candidate's *observed* score will be on a test, given the *true* score (the inverse to the popular use described in the literature). A detailed explanation underpinning the correct application is outside the scope of this review. Harvill did conclude, however, that it is reasonable to use the SEM to estimate the true score band around an observed test score *if* the test has a reasonably high reliability coefficient and the observed score is not an extreme outlier from the mean of the observed test scores (Harvill 1991:186).

As discussed, some authors argue that the SEM is a better reporting index of the reliability and *quality of a test*, as compared to the reliability coefficient, since it controls for the variance of test scores and is therefore a more *test-specific* instrument, reflecting the accuracy of a test's measuring capacity (Tighe *et al.* 2010:8).

The use of the SEM has also been used to help educators in a new Australian medical school review and refine their summative examination system and standard setting strategy (Hays, Gupta & Veitch 2008:814). They used the SEM to define the borderline performance band around the pass mark and three lower performance bands in SEM intervals to -3 SEM. Subsequently, they evaluated the re-sit assessment performance of students in each of the bands and found that it predicted future performance to a fair extent and it informed their assessment regulations in a defensible manner (Hays *et al.* 2008:814).

In summary, the reliability of test scores is inherently part of their validity as well. The reproducibility of the results either adds to, or threatens, the validity of the inferences made from the test scores. If test scores are not reliable, given the stakes of the test, they cannot be deemed valid. Essentially, it means that the 'noise' levels are unacceptably high and the 'signal' cannot be heard, so the message or meaning contained in the test scores (the validity) is not trustworthy enough to make important decisions based on them.

2.2.5 The concept of standard setting or assessment calibration

The preceding discussion makes it clear that high quality decision making (i.e. with a high level of accuracy and certainty) regarding mastery of a subject and academic progression (pass/fail) requires high quality performance data that is valid and reliable. The next step is to ensure the appropriate 'calibration' of the performance assessment results (the scores or marks produced by the test) is done in a fair, defensible and explicable fashion.

The term 'calibration' as used in this context refers to the process of applying human judgment to test marks, by way of a chosen standard setting method, in order to assign meaning to them in relation to the difficulty of the assessment (Van der Vleuten 2010:174). This is the process and purpose of standard setting – to give meaning to the assessment results in terms of the difficulty of the assessment (Van der Vleuten 2010:175).

The importance of standard setting is easily explained using a simple analogy. If a student sits a test, which contains 100 test items, and scores 63/100 (63%) for the test, should the student be deemed competent on the assessed domain and passed? In most South African universities or CMSA Colleges the answer would be: "Yes, since the student obtained a score of more than 50%". However, this pass judgement might seem inappropriate (or a false positive judgement) if 145 students wrote the same test and the class average was 87%. This result would suggest that the test was easy for the cohort of students and hence a score of 63% should be viewed in a different light. Conversely, if the same student scored 63%, but this time the class average of the 145 students was 37%, this would suggest that a score of 63% is an indicator of excellent performance. The point is that a score of 63% on the test does not carry any meaning on its own, but needs to be contextualised (or calibrated) in a methodical manner to

give it meaning and enable educators to make valid deductions about the competence of candidates based on the test results (Downing 2003b:830; Kane 1994:425-426).

Another way to conceptualise and explain standard setting is to compare the process with a similar process – the *calibration* of laboratory tests. In clinical practice, the clinician is confronted with a patient who, after initial assessment, might require a tissue sample (e.g. a blood sample) of the patient to be sent to a laboratory to undergo a specific test(s) to investigate for a particular illness. The result of the test(s) is then sent to the clinician to aid him or her making a decision regarding the patient's management. The integrity of this process is critical to making the correct decision. Possibilities of errors in the process include the wrong patient's sample is sent, the sample is sent in the wrong tube for the intended test, or the laboratory equipment is malfunctioning or uncalibrated, producing results that cannot be trusted. Quality control in the health services is needed to prevent errors and ensure that calibrated (accurate, valid and reliable) data reaches the clinician. Only then is the clinician able to make appropriate decisions, based on the test results.

In medical education assessment, essentially the same process is followed, leading to educators making decisions about the learning 'health' of an individual learner. The results emanating from the educational test(s) must to be calibrated (Van der Vleuten 2010:174) in terms of difficulty for the cohort under evaluation, to enable educators to make accurate decisions about who is educationally 'healthy enough' to progress in the programme or course, and who is not. Figure 2.4 schematically illustrates the assessment calibration (standard setting) concept as it relates to clinical practice.

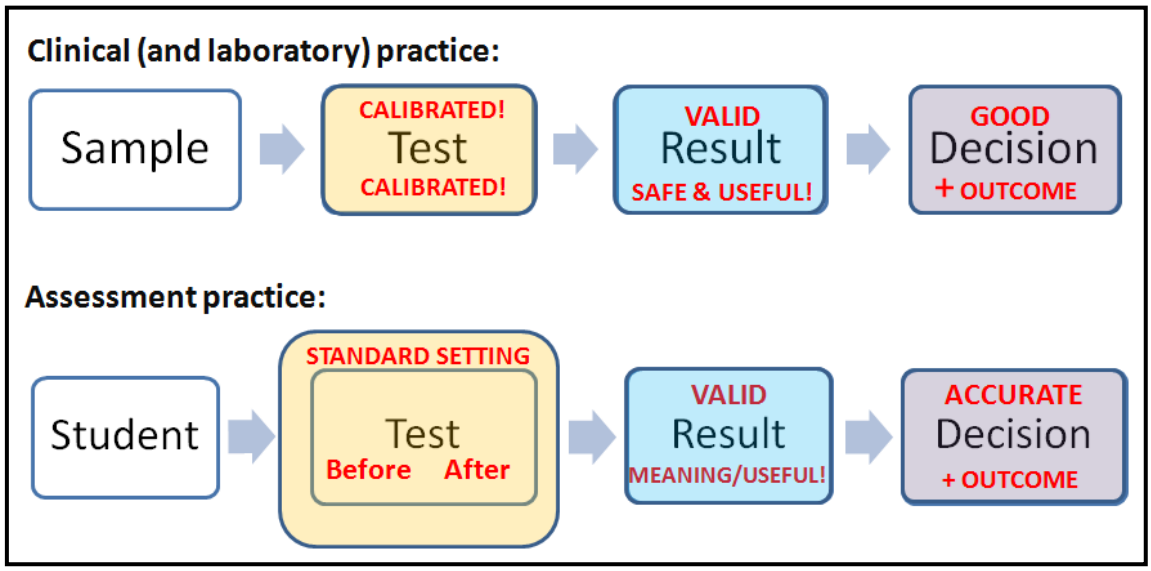


FIGURE 2.4: THE ASSESSMENT CALIBRATION PROCESS
[Compiled by the researcher, SCHOEMAN 2014]

This process of calibrating the educational tests or assessments can occur either prospectively before the test by a panel of judges evaluating the individual items in terms of their difficulty (test-centred methods) or retrospectively after the test by evaluating the psychometrics and analytical statistics of the produced results (examinee-centred methods) or a combination of both (some modified test-centred methods and some compromise methods for example the Hofstee method). Either way, human judgment needs to be applied to the test to give meaning to its results and to make subsequent fair and accurate decisions about the performances of the students.

The above examples made the point that objective written assessments of medical knowledge, comprising of well-written clinical case-based, single best-answer MCQ items or extended matching items (EMIs), might effectively sample from the curriculum and test clinical problem-solving on a wide range of problems posed to the candidates, (valid test) and do so in a reproducible manner (reliable test), but the inherent *difficulty* or standard of the test for the cohort under assessment remains mostly unknown to the test administrators. Calibrating the assessment difficulty (determining its standard), using a formal standard setting method, provides meaning to test results, which allow for their meaningful interpretation and for making appropriate pass/fail decisions (Hansen, Lyon, Heh & Zigmond 2013:301).

From the preceding text, the concept of standard setting is, therefore, clearly rooted in the question: "How much is enough?" (Cusimano 1996:s112; Downing *et al.* 2006:51; Livingston & Zieky 1989:121). In other words, what proportion of the learning outcomes/objectives need to be achieved, as measured by the assessment instrument, in order to be deemed a passing candidate? Many authors, books and review papers written over the past 40 years have discussed this issue at great length in both the general and medical education literature (Bandaranayake 2008; Barman 2008; Cizek, Bunch & Koons 2004; Cusimano 1996; Hambleton, Jaeger, Plake & Mills 2000; Hambleton & Pitoniak 2006; Livingston & Zieky 1982).

Kane (1994) describes the relationship between the pass mark of an assessment and the abstract construct of the performance standard as follows: "the performance standard is the conceptual version of the desired level of competence, and the pass mark is the operational version of the desired level of competence" (Kane 1994:426). Standard setting is therefore defined as the process by which educators convert or

operationalise the abstract construct of "the performance standard" into an operational or actual cut-point on the continuum of an assessment scoring (or result) scale, commonly referred to as the pass mark, passing score or cut-score (Bandaranayake 2008:836; Barman 2008:957; Hambleton & Pitoniak 2006:435; Norcini 2003:464). When examinees achieve a test score on or above the pass mark, they are classified as competent (acceptable performance) to pass the particular assessment. Conversely, if a candidate scores less than the pass mark, the performance is classified as unacceptable and the candidate fails the assessment (Cusimano 1996:s112).

As Kane (1994:433) explains further "the pass mark is the particular point on the score scale that is used operationally to make decisions about examinees and the performance standard is a conceptual boundary between acceptable and unacceptable levels of achievement. The pass mark is a number, and the performance standard is a construct." Standard setting methods is, therefore, tasked with converting the expected, abstract performance standard into a numerical pass mark on the scale of the test scores.

2.2.6 The purpose of standard setting in medical education

The purpose of standard setting is closely aligned with one of the main purposes of assessment in education (Purpose 1 as explained in section 2.2.1 above) – making judgement decisions (pass or fail) about examinees' mastery of the learning outcomes or leaning objectives as measured by the relevant assessment instruments (Wass *et al.* 2001:945). This matter of making judgement decisions about examinees is of critical importance in professional education, such as medical education, because licensed candidates, who have been certified as being competent by the assessment system, can now practice their profession on the public (the patients in the case of medical education).

The decision makers or 'gatekeepers' in undergraduate and postgraduate medical education are the academic staff members of medical schools, national examining and certification bodies and, indirectly, the policy makers in the regulatory bodies, which govern the institutions and organisations providing the medical education and their respective assessments of clinical and professional competence. Collectively, these individuals are henceforth referred to as 'educators' in this thesis.

Since the issue of making pass/fail decisions about the performance of candidates on assessments is at the heart of standard setting, it is important to discuss the implications of these decision-making processes. Naturally, educators need to carefully consider these very important judgement decisions, given its wide ranging impact both on the individual trainee and the public.

This places medical educators in a 'gatekeeping' role between, on the one side, aspiring new clinicians and on the other side, the public who are the recipients of healthcare. The educators need to be fair to the students and not fail them without good justification, but also protect the public and not pass the students who might place the public at risk due to a lack of professional competence (Bandaranayake 2008:837). Therefore, medical educators have an important social responsibility to both the public and the learners within the medical education and training process (Bandaranayake 2008:837). Understanding this *social accountability* role of standard setting in medical education is important and is further discussed in section 2.2.11 of this chapter.

However, what is fairness to students with regard to testing (assessment)? Crocker and Zieky state: "it is impossible to set fair standards on unfair assessments, and that it is impossible to set valid standards on invalid assessments" followed by "The fairness of standards depends first of all on the fairness of assessments" (Crocker & Zieky 1995:19). Zieky (2002:1) provide a different point of view to assessment fairness and endorse the definition of the Educational Testing Service in the USA regarding test fairness in high-stakes tests as: "Fairness requires that construct-irrelevant personal characteristics of test takers have no appreciable effect on test results or their interpretation" (Educational Testing Service 2002:17).

This is not an easy task for educators and it clearly adds to their already significant responsibilities. The assessment instruments and standard setting methods used by educators to make decisions at the various levels of education and training, therefore, need to be fair and robust, yet affordable and feasible in the local context (Van der Vleuten 1996:41,61-62). These essential characteristics of standard setting should be incorporated into the utility framework used to evaluate standard setting methods.

2.2.7 The principles of standard setting

The key principles of standard setting were briefly outlined in Chapter 1. An in-depth discussion and review of the relevant literature is provided in this section of the chapter.

2.2.7.1 *All based on human judgement*

In a review paper on standard setting, Hambleton and Pitoniak (2006) explain that all aspects of an education system are designed and constructed by human judgement. This includes the design, content, length of the curriculum, which resources are prescribed, the teaching methods, as well as the assessment system, including how the pass standard will be determined. This is a critical principle in understanding the nature of *all* standard setting methods and has been extensively discussed by a large number of publications in the general literature (Cizek 2013:7; Cizek *et al.* 2004:31; Hambleton & Pitoniak 2006:435; Kane 1994:425; 1998:130; Livingston & Zieky 1982:12; Zieky 1995:30), and the medical education literature (Bandaranayake 2008:837; Barman 2008:959; Cusimano 1996:112; Downing *et al.* 2006:51; Norcini 2003:464; Schuwirth & Van der Vleuten 2010:204). The ways in which these judgements are made vary, depending on the particular method used to set the passing standard of an assessment process.

In a recent workshop presented to academic staff at the University of the Free State, as part of a mini-conference on assessment in medical education, Steward Petersen explained that making judgements about the performance of students or candidates in written assessments has always been an integral part of the process of progression in medical education (Petersen 2013). Until the 1970s, when the prominence of MCQs and short-answer questions (SAQs) started to increase dramatically, the 'grading of essays' was an important part of written assessments. This process of grading or marking the essays was mostly based on the overall view or opinion of the marker, regarding the student's mastery or understanding of the knowledge and/or concept assessed in the particular essay. The marker then calibrated the student's essay using the difficulty or complexity of the essay question, and the educational level of the course and their own internal standards as an expert in the field. Based on this intuitive process, the essay was awarded a particular grade. Standard setting, in this context, was embedded within the actual marking process (Petersen 2013).

In many other higher education contexts the marking of written constructed-response criterion-referenced assessments is done holistically and qualitatively without the use of marks and is generally referred to as the *grading* of assessments (Sadler 2005:177). Sadler (2005:179-186) describes four grading models in constructed-response criterion-referenced assessment, and these vary with regard to the assessment context and the extent of the qualitative versus quantitative nature of the grading processes.

As previously explained in the chapter, the use of essays has decreased substantially in medical education assessment over the past 40 years, and has largely been replaced with predominantly MCQs and SAQs, mostly for psychometric and efficiency reasons. This move to using more objective and reliable assessment instruments, that can efficiently cater for large numbers of students in one sitting, has meant the loss of the expert human marker, in providing a judgement on each student's assessment result in relation to the difficulty of the actual test. This issue, together with 'grade inflation' in secondary school education, leading to students receiving grades higher than their actual ability, were two of the main driving forces behind the development and expansion of the standard setting field in the American education system and educational literature in the 1970s and 1980s (Cizek 2013:6-8).

Most of the more established standard setting methods, such as Angoff (Angoff 1971), Ebel (Ebel 1972) or Nedelsky (Nedelsky 1954) centre around a panel of expert judges deciding how a hypothetical borderline candidate (i.e. a minimally- or just-competent) would fare in a test. This introduces a measure of bias into the decision, since each judge has his or her own ideas and perspectives about what constitutes borderline ability and how a hypothetical borderline student or group of a 100 such students would answer each item in the test.

It is these two aspects of panel-based test-centred (PBTC) methods – the conceptualisation of the hypothetical borderline student and estimating (predicting) how this borderline student will actually perform on the test items, that have been the source of serious concern and criticism of these methods (Boursicot & Roberts 2006:85; Hansen *et al.* 2013:301; Wayne, Fudala, Butter, Siddall, Feinglass, Wade & McGaghie 2005:S65). Berk reported that prominent national testing organisations in the USA went as far as to call these methods "fundamentally and conceptually flawed" and called for new approaches to be found (Berk 1996:216).

This issue around conceptualising 'the borderline candidate in the actual cohort under assessment' is a well-documented conundrum in the literature (*cf.* Barman 2008:959; *cf.* Hambleton & Pitoniak 2006:441). Firstly, to define and reach consensus about what actually constitutes a 'borderline candidate' is a difficult task for experienced expert judges, with many years of experience in the field (Bandaranayake 2008:841).

Secondly, judges have voiced their concern about their tasks in standard setting panels – to estimate the probability that a borderline student will answer the item correctly (selected-response items), or what a borderline student will score on each item (constructed-response or performance-based items) and then translating their predictions into numerical values in a PBTC-method, such as the Angoff method (Boursicot & Roberts 2006:85; Impara & Plake 1997:354). To make this difficult task easier, some judges (of their own accord) have moved away from the consensus derived characteristics of a hypothetical borderline student, to a well-known *real* student who is a borderline performer in their perspective. They then ask themselves how this real borderline student would perform on the test items (Impara & Plake 1997:355).

Thirdly, the unrealistically high passing scores generated by many panels, of experienced experts in the field, have been the main driver behind the development of one of the most common modification of PBTC methods - the "reality check" (Norcini, Shea & Kanya 1988:63; Ricker 2006:55). The modification requires that the panel be provided with real performance data on the same test, or a very similar one, from previous cohorts of candidates, to enable them to 're-calibrate' their standard setting gauges towards reaching a more *realistic* judgement and expectation of the cohort under assessment.

This modification, classically described as part of the many *modified* Angoff method variants (Hambleton & Pitoniak 2006:441) is discussed in more detail in the section about the Angoff method, later in this chapter. However, it suffices to say at this stage that, the reality check modification is probably the best indicator of how hard it is for examiners to set consistent (reliable), accurate and realistic (valid) pass standards on their own, with a purist strategy. Indeed, according to Berk (1996:216), the only consensus opinion, that currently exists within the standard setting community, is that different panels may reach different pass marks using the same method, and that

different methods will probably yield different pass marks for the same test and cohort of candidates under assessment.

Despite all the concerns about PBTC methods, they remain the most widely used standard setting methods in all forms of education, including medical education (Bandaranayake 2008:959; Clauser, Mee, Baldwin, Margolis & Dillon 2009b:390). The reason for this, ironically, is because the pass mark is based on the consensus judgement of a panel of experts in the field, who have reviewed the test items.

From the wealth of literature on PBTC methods, it seems that the magical ingredient is a 'panel of experts in the field' making the judgements. They are respected by colleagues in the medical profession, the courts and the test candidates, and so *their* judgements are perceived to be valid and trustworthy. It is no surprise then, that a large section of the literature on standard setting deals exclusively with the required qualifications, selection, training and evaluation of panellists (Hambleton & Pitoniak 2006:451-456; Livingston & Zieky 1982).

Since all standard setting methods are based on human judgement, and test administrators assemble the best available panel of experts in the field, should attempts to improve standard setting not rather focus on what panels are asked to do? Visualising a hypothetical borderline student and estimating his/her performance on an upcoming test is fraught with controversies, both cognitively and psychometrically (Clauser *et al.* 2009b:390-391). Attempts to address these controversies and reduce their negative impact have made PBTC standard setting processes longer, more time consuming and more complex, with multiple modifications in the form of discussions, reality checks and iterations on judgements, which drive up costs and erode the long-term feasibility and sustainability of using these methods, especially in resource-limited settings, such as South Africa. From this local context perspective, this might have been a contributing factor why, despite the importance of standard setting being widely emphasised in many policy documents relating to the medical education since the birth of the new South Africa in 1994 (Hift & Burch 2003:76; Lindgren, Ahn, Alwan, Cassimatis, Jacobs, Karle, Kloiber, Van Lerberghe, Patricio, Pulido, Sood & Weggemans 2012:25), procedural standard setting is at present so limited in the South African general and medical education landscape (Pitoniak & Yeld 2013:23; Schoeman 2011:2). This challenge is discussed further in the contextualisation part of this chapter.

Perhaps the time has come for us to ask and focus the attention of our scarce resource, the panel of experts in the field, on a different aspect of judgemental decision making: the aspect of evaluating the actual performance data of the cohort who have sat for the written assessment and for whom we need to set a pass mark.

Given the limitations of the PBTC methods, it is appropriate that there has been a shift towards developing standard setting methods which ask panels of expert judges in the field to evaluate performance data. This strategy is not new, for example the Hofstee method was described in the early 1980s (Hofstee 1983). These methods, discussed in more depth later in this chapter, are referred to as 'compromise methods'. They ask expert panels to make judgements, usually before the administration of the test (and thus set criteria), based on the test items, about *inter alia* minimum and maximum pass/fail rates, minimum and maximum pass marks and the confidence of the panels in their decisions. These judgements are then used as a framework or grid onto which the test performance data of the examinees are plotted and decisions about the pass mark and passing (or failing) rates are made (Beuk 1984; De Gruijter 1985; Hambleton & Pitoniak 2006; Hofstee 1983).

Finally, expert panels can also be asked to give input towards developing a model to 'bench mark' the difficulty and pass/fail score using test data. The Cohen method (Cohen-Schotanus & Van der Vleuten 2010:157), discussed in depth later in the chapter, focuses the attention of the expert panel on two aspects: i) which point of reference should be used as an indicator of test difficulty, and ii) what percentage of the reference point will be used as the absolute passing score (pass mark)? The review of previous performance data on similar tests and student cohorts can aid the panel's judgement decisions, (Cohen-Schotanus & Van der Vleuten 2010:159) when addressing the two aspects mentioned.

2.2.7.2 *There is no 'GOLD standard' method*

The second important and consistent outcome from the literature review on standard setting in all spheres of education, including medical education, is that there is no consensus on which method is considered the best method for all assessments instruments. In short, there is no 'gold standard' method (Barman 2008:958; Ben-David 2000:120; Cohen-Schotanus & Van der Vleuten 2010:154; Downing *et al.*

2006:50; Norcini 2003:464; Schuwirth & Van der Vleuten 2010:204; Taylor 2011:e678; Van der Vleuten 2010:174). No one size fits all, and different contexts and available resources play a significant role in deciding, which methods should be used (Cohen-Schotanus & Van der Vleuten 2010:154). Given this reality, there are a variety of standard setting methods in use that can be selected for a particular assessment instrument, for example written MCQs versus an Objectively Structured Clinical Examination (OSCE) (Schuwirth & Van der Vleuten 2010:205). A classification of standard setting methods, with examples of commonly used methods, is presented later in this chapter.

2.2.7.3 *The arbitrary nature of setting standards*

Since there is no 'gold standard' standard setting method and all methods are based to a greater or lesser extent on human judgements, it means that all pass standards set for any assessment are arbitrary in nature (Barman 2008:959; Norcini 2003:464; Schuwirth & Van der Vleuten 2010:205).

Although this principle sounds alarming for educators in the medical education community, it does not need to be. As Schuwirth and Van der Vleuten (2010:205) explain, although all pass standards are arbitrary in nature, because they are constituted on human judgement, they are *never* without huge importance and value and, as such, they should always be i) explicable, ii) defensible, and iii) stable between cohorts of students.

Therefore, concerns arise when an institution or examination body *cannot* explain to stakeholders i) how the pass mark was derived (i.e. no formal standard setting process or methods was used), or the method used is too complex and difficult to understand (i.e. not explicable); ii) why the particular pass mark is valid and credible, or demonstrate that proper effort was put into determining the pass mark (i.e. not defensible); and iii) why the pass mark or cut-score varies widely between different cohorts of students from year to year (i.e. not stable) (Schuwirth & Van der Vleuten 2010:205).

Van der Vleuten (2010:174) also makes the point that unless a particular standard setting method explicitly takes account of the *difficulty* of the test, that particular method is not defensible. He explains that in some of his unpublished research at

Maastricht University on variance component estimation, conducted on seven cohorts over four years of preclinical students, he noted that observed test score variance associated with test difficulty was *far greater* than the variance observed between cohorts of students (Van der Vleuten 2010:174). This paper makes a strong argument for incorporating test difficulty in standard setting processes.

2.2.8 Classification of standard setting methods

The classification of standard setting methods has undergone multiple revisions and modifications over the years (Cusimano 1996S112). A full historical overview of these classification developments falls outside the scope of this review. Instead, a contemporary classification is provided, based on the literature and *agreed* principles of standard setting.

The notion of 'agreement' in standard setting is rare, and has to be understood in the context of all the challenges educators face, when setting pass standards for different tests in different contexts. There is, however, some common ground and consensus has been reached in some aspects of standard setting. These are important aspects to consider when classifying different methods, because they delineate the methodological differences between the methods, which make a classification possible. Figure 2.5 outlines the basic classification of standard setting methods.

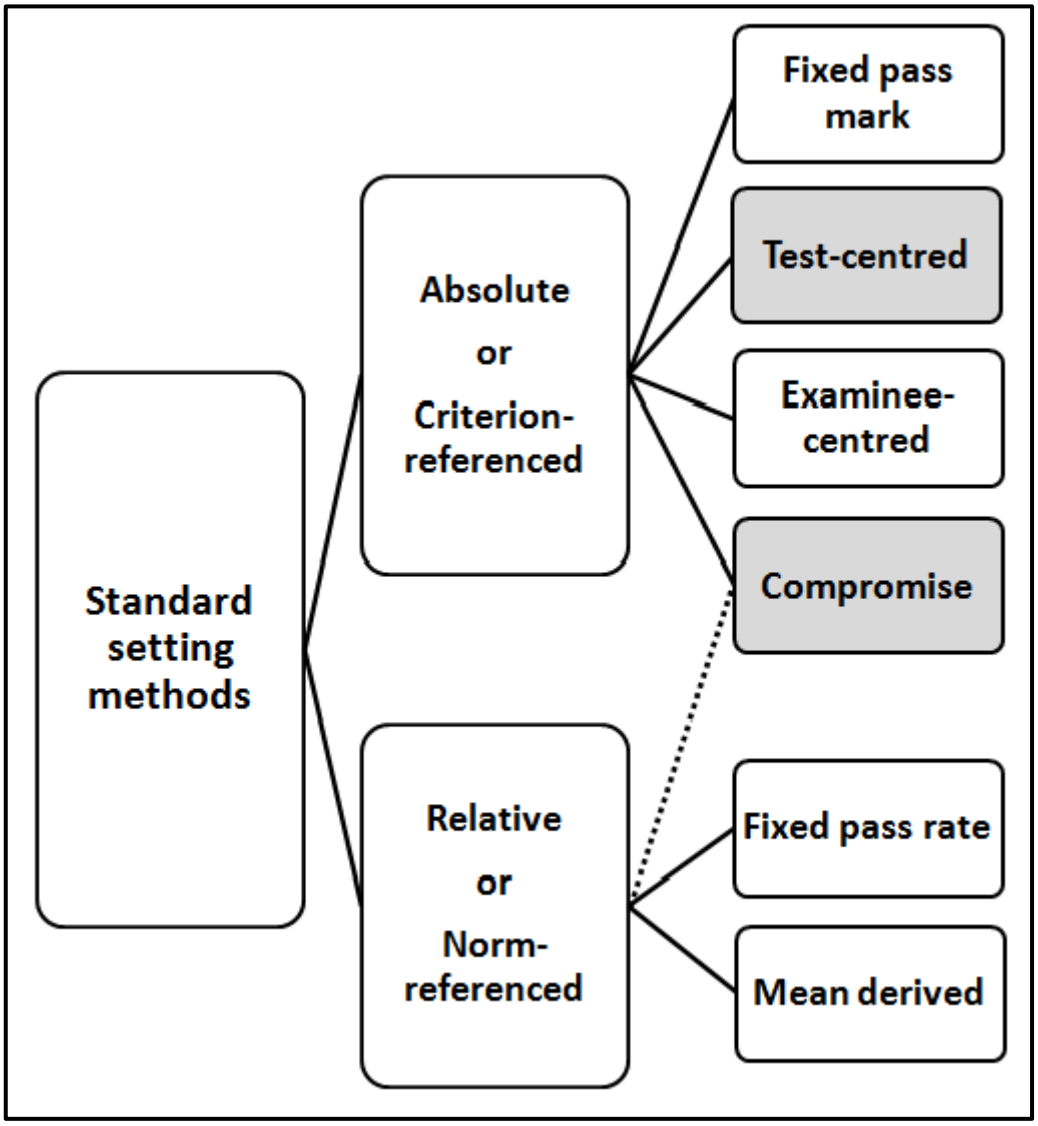


Figure 2.5: CLASSIFICATION OF STANDARD SETTING METHODS [Compiled by the researcher, SCHOEMAN 2014]

Absolute/Criterion-referenced or *Relative/Norm-referenced methods* is the most basic level of classifying standard setting methods. The differences between these two broad types are explained in more detail in the next section, but a brief summary, including their respective sub-divisions (*cf.* Figure 2.5) and some examples are provided here.

Absolute / Criterion-referenced methods

- In these methods, the translation or conversion of the conceptual performance standard into a numerical pass mark (score) on the observed score scale of the test is achieved by using panels of judges to set the *absolute* pass criterion (Cizek *et al.* 2004:32; Kane 1998:130).

- The decision is made exclusively by a panel of judges (test-centred and examinee-centred methods) or partly with additional data inputs (compromise methods).
- The concept of the 'borderline' or 'minimally competent' student is a central theme in the test-centred and examinee-centred methods.
- All the candidates can pass the test if all are deemed competent.
- *Fixed pass mark:* the traditional fixed 50 – 60% pass mark, depending on context or country (Schoeman 2011:2)
- *Test-centred methods* include: Nedelsky (Nedelsky 1954), Angoff (Angoff 1971), Ebel (Ebel 1972), Jaeger (Jaeger 1989), Bookmark (Lewis, Mitzel & Green 1996)
- *Examinee-centred methods* include: Borderline group (Wilkinson *et al.* 2001), Borderline Regression (Wood *et al.* 2006), Contrasting Groups (Livingston & Zieky 1982)
- *Compromise methods* include: Hofstee (Hofstee 1983), Beuk (Beuk 1984), De Gruijter (De Gruijter 1985), Cohen (Cohen-Schotanus & Van der Vleuten 2010), Objective Borderline (Shulruf, Turner, Poole & Wilkinson 2013)

Relative / Norm-referenced methods

- In these methods the translation or conversion of the conceptual performance standard into a numerical pass mark (score) on the observed score scale of the test (Cizek *et al.* 2004:32; Kane 1998:130) is not explicitly derived. The assumption is made that a certain percentage of candidates have achieved competence and the rest have not.
- Not all the candidates can pass the test (even if all are possibly competent and the assumption above is incorrect). The test scores of the candidates is rank ordered from the highest to the lowest score and only a proportion of the candidates are deemed to have passed the test.
- Not advised for use in competency assessment, since the standard set is not directly linked to the competency of the candidates.
- *Fixed pass rate* methods – e.g. the top 35% of candidates pass (McManus, Mollon, Duke & Vale 2005:2; Zieky 1995:53)
- *Mean-derived* methods – e.g. Wijnen method, where the pass mark is set as two times the standard error of measurement (SEM) below the mean score (Wijnen 1971).

2.2.8.1 *Absolute methods or criterion-referenced methods*

The term 'absolute' implies that the pass standard is set by a panel of judges, who are *blinded* to the actual performance data of the cohort they are setting the pass standard for, and therefore, focus exclusively on the content of the test (test-centred methods) or the performance prompts (examinee-centred methods). This is the *purist* view of criterion-referenced methods according to Yudkowsky and Downing (2009:L1826).

Unfortunately, in reality, pure absolute standards rarely turn out to be realistic, acceptable, or useful in the real world of medical education (Yudkowsky & Downing 2009:L1826). Judges usually have too high expectations of borderline candidates and in the absence of real performance data to provide a 'reality check', they set pass marks that would yield unacceptably high failure rates (Cizek 2001:391; Clauser *et al.* 2009b; Kellow & Willson 2008:19). Any absolute/criterion-referenced method, that engages with actual performance data to help judges provide more realistic ratings, essentially becomes a compromise method (*cf.* discussion on compromised methods later in this chapter).

A more practical and accurate definition of an 'absolute' pass mark is, that all candidates who achieve it will pass the test and it is set in such a manner that all candidates could potentially achieve it. This means it is not set in a way that a certain amount or proportion of candidates must fail the test, irrespective of their performance, which is the case with relative/norm-referenced methods.

Fixed pass mark

However, it is important to note that not all absolute/criterion-referenced methods produce defensible pass marks. The arbitrary, pre-set, fixed pass mark, usually set by institutions at 50% to 60%, is also an example of an absolute pass mark (Schoeman 2011:2).

Searle (2000:363) and Bhandary (2011:3), however, argue strongly that although it is easy to set a pass mark of 50%, it is not a fair measure to determine who is competent or not and that it is not transparent nor is it defensible. The concerns with this method is that there is *no* link to the standard of the test and it is completely insensitive to the *difficulty* of the test, (Schoeman 2011:2; Van der Vleuten 2010:175),

which means it has an unknown relationship with competence (Searle 2000:366). This method is mostly based on tradition (Zieky 1995:33) and as a result, it is widely used in medical education institutions (Van der Vleuten 2010:175), including South Africa, at all levels of undergraduate and postgraduate medical education (Schoeman 2011:2).

For the reasons just described, this method is regarded as indefensible and, like norm-referenced methods, may lead to increased false positive or false negative decisions about the progression of candidates who sat the test (Barman 2008:959; Cusimano 1996:S117).

Test-centred methods

The concept of the 'borderline' student and the challenges it brings to PBTC standard setting methods, have been discussed in the thesis and further discussion is not needed.

Examples of PBTC methods were provided in the previous section. The Angoff method and its common modifications are discussed extensively in the next section of this chapter, because it was the PBTC method used in this study. Short descriptions of other commonly used test-centred methods, highlighting their unique features, are provided below.

Ebel's method (Ebel 1972; *cf.* Hambleton & Pitoniak 2006:442) is a widely used method, which has been extensively researched over many years. It is similar to the Angoff method and is based on the same principle of item-by-item reviewed, panel-based, test-centred methods, i.e. the probability of a hypothetical borderline candidate, from the assessed cohort, correctly answering individual test items. However, unlike the Angoff method the judges must classify each item's *difficulty* for a borderline student as easy, moderate or hard (3 options).

In addition to judging the difficulty of each test item, panellists must also judge the *relevance* of each item to the cohort's learning goals as essential, important, acceptable or questionable (4 options, different categories have been used by educators using this method). A 3x4 matrix or grid is then constructed with 12 boxes where the judges must then *predict the proportion* of 100 borderline students who will answer the item correctly, (for MCQ-type items) or what they are likely to score (for

constructed-response or performance items) for each of the 12 possible item difficulty/relevance options. The pass mark is then calculated by multiplying the consensus number of items per box, with the consensus proportion of borderline candidates who are predicted to answer that type of item correctly. The mean of the 12 boxes is the pass mark for the test.

The Ebel method can be used for written or performance assessments. More judgements per test item are needed (difficulty and relevance) than with Angoff. Items are essentially placed in one of 12 boxes and the probability of a borderline candidate's performance is judged per box option (usually 12) and not per item, as in the case with Angoff. Modifications, similar to the Angoff modifications – multiple rounds of discussion; providing real performance data on the items of previous similar cohorts and 'impact' data showing how many pass/fail candidates the panel's current judgements will yield.

Nedelsky's method (*cf.* Hambleton & Pitoniak 2006:442; Nedelsky 1954) was published in 1954 and is probably the oldest formal assessment standard setting method described. It was designed specifically for use with MCQ tests and its use is limited to that format. It works on a simple strategy of asking a panel of judges to review each MCQ test item and decide collectively which of the MCQ *options* a hypothetical borderline candidate, from the assessed cohort, will identify as incorrect. If a particular MCQ item has four possible options, and the panel agrees that a hypothetical borderline candidate will eliminate two of the three wrong options, with one wrong option and the correct option remaining as plausible for the borderline candidate, the pass mark for the items is calculated as 2/4 or 0.5. It represents the probability of random guessing by a borderline candidate on the item. The pass mark of the test is calculated by summating the probabilities per item.

The Nedelsky method is widely used and has been extensively researched, but its use is declining because emerging research suggests that it sets the standard too low (mostly in comparison to Angoff) (Chinn & Hertz 2002:3; Hambleton & Pitoniak 2006:442). This is explained by the limited and fixed probability options the method provides, depending on the number of MCQ item options. Most items commonly end with 0.5 or lower probabilities, which results in low pass standards being set for the test.

Jeager's method (*cf.* Hambleton & Pitoniak 2006:442; Jaeger 1989) is different from most of the commonly used item-by-item reviewed, panel-based, test-centred methods, in as much as how panels are selected and the judgement task given to the panellists. Usually PBTC standard setting panels consist mainly of 'expert discipline-specific gatekeepers' (the educators – faculty staff members, examiners of the College, etc.) and sometimes recent graduates from the course (Verhoeven, Van der Steeg, Scherpbier, Muijtjens, Verwijnen & Van der Vleuten 1999:833). In the Jaeger method, the panel should consist of representatives from *all the stakeholder groups* on whom the outcome of the test will have an effect and, therefore, for whom the particular pass standard being set is important. In the case of medical education, it would be educators, regulatory stakeholders, employers, candidates and the lay public.

The task for Jaeger panellists is to provide a Yes/No judgement of whether every passing candidate from the assessment *should* be able to answer any given item correctly. It is, therefore, not a probability judgement (would), but a *value* judgement (should).

The Jaeger method was originally described for use in secondary-school graduating examinations in the USA. It is not widely used in medical education, given the practical difficulty in assembling an appropriate panel and due to the unrealistically high pass standards it would probably set. The only example of a modified-Jaeger method (traditional expert panel used) in the medical education assessment literature was in the MCRP (UK) assessments from 2002 up to 2008, where panellists were asked to make judgements on MCQ items in a value-laden manner: *should* a borderline candidate from the assessed cohort be successful on this item? (McManus, Chis, Fox, Waller & Tang 2014:3). The authors referred to it as an Angoff procedure, and controlled for high failure rates by adding a Hofstee component for setting the final pass mark.

The *Bookmark method* (Karantonis & Sireci 2006; Lewis *et al.* 1996; Lypson, Downing, Gruppen & Yudkowsky 2013:582) is based on the concept of ordering the test items of an assessment event in ascending order of difficulty for the cohort, i.e. starting from the easiest item and ending with the hardest one. Determining item difficulty was originally done as using Item Response Theory (IRT) and item mapping or construct maps (Wyse 2013), but some authors have proposed using the conceptually easier

Classical Test Theory (CTT) system of proportion correct answers on the item in the test (the PC-value) (Buckendahl, Smith, Impara & Plake 2002).

The panel of judges then need to agree up to which item, from easy to most difficult, 67% of borderline candidates would give a correct answer. The bookmark is placed just after the 'last item correct' position. The pass mark is then calculated as the 'last correct item' number divided by total amount of items in the test. The use of 67% of borderline candidates being successful is controversial in the literature and many authors have questioned it (Hambleton & Pitoniak 2006:443). The explanation offered is that 50% would constitute 'uncertainty' about whether a borderline candidate would be successful and values greater than 67% could lead to unrealistically high pass standards being set. A literature review published by Karantonis and Sireci (2006:8) found that the majority of the evidence supported the use of 67% for the method.

The bookmark method has been described for use with written (MCQ) and performance assessments. It has also been used in medical education and its performance has been favourably compared to the Angoff method (Buckendahl *et al.* 2002; Cetin & Gelbal 2013; Peterson, Schulz & Engelhard (Jnr.) 2011).

Examinee-based methods

Examinee-based methods are predominantly used for performance-based assessments of clinical or practical competencies such as the Objectively Structured Clinical Examination (OSCE) (McKinley & Norcini 2014:98-99). The reason for the specific focus of examinee-based methods is mainly due to their inherent design. They require expert judges to make competency judgements based on the observed performance of examinees performing a given task(s) (McKinley & Norcini 2014:101).

Performance-based assessment methods were not studied as part of this research project and hence, examinee-based standard setting methods will not be discussed at length. Important and prevalent aspects are highlighted and referenced examples are provided.

The concept of the 'borderline' candidate is still used for the *Borderline Group* (BG) (Wilkinson *et al.* 2001) and *Borderline Regression* (BR) (Wood *et al.* 2006) methods. However, the borderline concept is not such a concern here, since it is not related to a *hypothetical* borderline candidate. Judges are required to recognise *actual* borderline

performance in the examinees they assess while they perform the required tasks in a performance-based assessment, for example in a physical examination OSCE station.

Examiners are typically asked to score the performance of the candidates in their station using a standardised item checklist (Norcini & McKinley 2007:240) or a more holistic rating-scale (Boulet, De Champlain & McKinley 2003:245) of the skills in the task, such as clinical communication. In addition, the examiners are asked to provide a global or overall opinion about the performance of the candidate at the end of the assessment event. This is usually done using a separate scale on the mark sheet. A classic example of the scale would be: fail, borderline, pass, well done. It is at this point that the BG and BR methods diverge in methodology in terms of calculating the pass mark for the station.

In BG method, the totalled checklist or rating-scale score for each candidate is plotted against his/her global impression rating. The median task checklist or rating-scale score of all the candidates who were globally rated as 'borderline' is then used as the station pass mark (McKinley & Norcini 2014:101). In the newer BR method, the task scores and global impression ratings are plotted for all the candidates who sat the assessment. A regression line is then drawn through the data and the point where this line crosses the 'borderline' group, is taken as the station pass mark. Therefore, in the BG method, only the borderline group's data are used, while all the data from all the candidates are used in the BR method. The overall examination's pass mark is derived from averaging the station pass marks.

The other useful aspect of the BG and BR methods is their efficiency. There are no lengthy panel meetings needed outside of the examination, since the judgements take place *during* the examination. It does, however, mean that expert judges must be used (to judge the skills) and they must be trained and properly briefed before the start of the performance assessment (McKinley & Norcini 2014:100).

The advantage of the BR method, over the BG method, is that in statistical terms, more data points are advantageous from a reliability perspective and hence the BR method can be used on smaller cohorts of candidates than the BG method (Wood *et al.* 2006:122). The BR method is at present probably regarded as the preferred standard setting method for OSCEs due to its favourable utility profile (Pell, Fuller, Homer & Roberts 2010:804).

In the BG and BR standard setting methods the examiners function predominantly independently to judge the examinees. However, in the *Contrasting-groups* method, they again function as a group, akin to the PBTC methods.

The *Contrasting-groups* method (Livingston & Zieky 1982) classifies a cohort of candidates into two 'contrasting' groups – competent and incompetent. This method can be used in both written and performance based assessments. It is labour intensive, since the performance of the candidates must be scored first (Downing *et al.* 2006:56) and thereafter submitted to this standard setting method for review in order to set the pass mark (Livingston & Zieky 1982:35). The strategy used to achieve this is by reviewing the performances of a sample or all of the candidates and then classify the performance as competent (pass) or incompetent (fail). A frequency plot is then drawn for each of the pass and fail groups, using their respective performance scores as assigned during the initial grading process. Where the pass and fail frequency plots intersect is usually taken as the pass mark on the original grading scale (McKinley & Norcini 2014:106)

Compromise methods

The last group of methods in the Absolute/criterion-referenced domain is the Compromise methods. These methods were developed to try and minimise the challenges faced when using *pure* absolute/criterion-referenced test-centred methods, i.e. unrealistically high pass marks marked by unacceptable high failure rates as discussed in the preceding text. These mostly centre around the limited ability of panellists to predict the performance of a hypothetical borderline candidate.

By contrast, the examinee-based absolute methods have a 'built-in' normative 'realism' since the actual performance of candidates is directly observed and assessed by judges. In addition, the best aspect of relative/norm-referenced methods is their sensitivity to the *difficulty* of the test (Cohen-Schotanus & Van der Vleuten 2010; De Gruijter 1985; Van der Vleuten 2010). This advantage of sensitivity to test difficulty is used by compromised methods to ensure that realism is central to the pass marks they set and why several authors have argued strongly for their use (Cusimano 1996:S116; Livingston & Zieky 1989:137; McManus *et al.* 2005:2; Van der Vleuten 2010:175). The incorporation of normative performance data or the addition of a norm-referenced component to the methodology of an absolute/criterion-referenced standard setting method is the essence of how compromise methods work.

Their basic methodology consists of panel-based discussions and the use of retrospective performance data (Cusimano 1996:S116). They all set absolute pass marks, but it is located in a realistic spectrum of the scoring scale. The 'reality-check' is incorporated into the methodology of compromise methods and not regarded as mere modification of an existing absolute method. The compromise is, therefore, between the isolated judgments of panellists (without the influence of normative data), based on their consensus views, and the performance data of the cohort. The two overarching reasons and motivation for the development of compromise methods are the need to set *realistic* pass marks and *practicability* of use (Cusimano 1996:S116).

The *Cohen method*, an example of a compromise method, is the major focus of this thesis and is discussed in detail in section 2.2.10 of this chapter. Other examples of compromise methods include the popular Hofstee method (Hofstee 1983), Beuk's method (Beuk 1984) and De Gruijter's method (De Gruijter 1985). A brief discussion of each is provided next.

In the *Hofstee method* (*cf.* Bandaranayake 2008:840; Hofstee 1983) the panel of judges review the entire test content and at the end are asked to provide four judgements: The maximum and minimum *pass marks* (C – cut score) for the particular test paper as well as the maximum and minimum *failure rates* (F). These panel-based consensus judgments (either by way of discussion or averaging the individual judgements) delineate the boundaries of acceptable outcomes of the test as defined by the 'gatekeepers' (the faculty/examiners/judges).

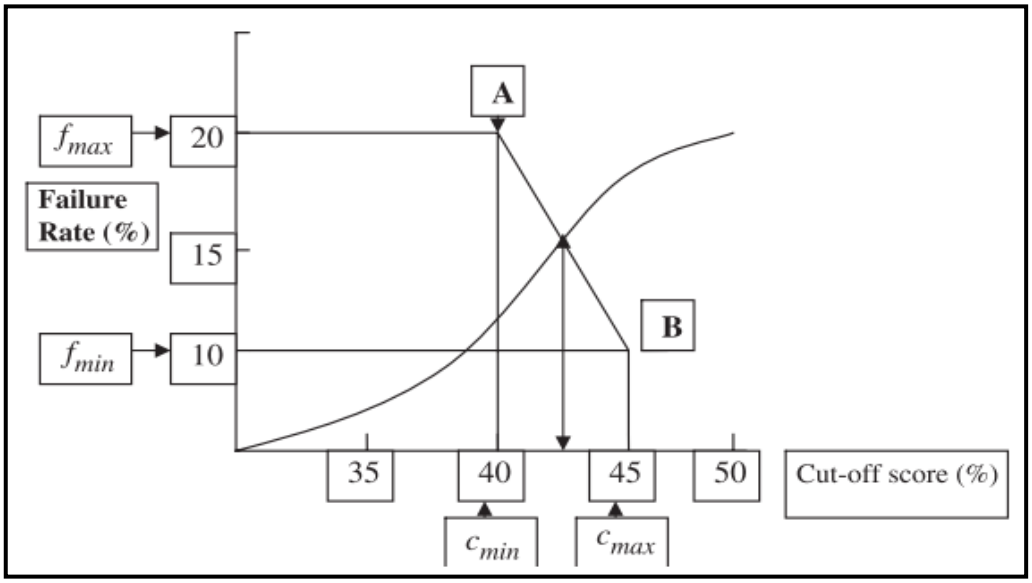


FIGURE 2.6: THE HOFSTEE METHOD
[Figure used from (Bandaranayake 2008:840)]

Figure 2.6 provides a practical example, where the maximum and minimum pass mark was decided as 45% and 40% respectively and the maximum and minimum failure rate was decided as 20% and 10% respectively for a particular assessment, usually written assessments. The panel's four collective judgements are then plotted on a graph with the score scale of the test on the X-axis, usually 0-100%, with the C_{max} and C_{min} pass marks and the percentage of students on the Y-axis, from 0-100%, with the F_{max} and F_{min} failure rates (*cf.* Figure 2.6). A cumulative frequency curve derived from the performance of the candidates is then added to the graph. The point where the line between the C_{max} and F_{max} points (A to B) intersects the cumulative frequency scores curve, is then used as the pass mark for the test. As can be seen in this example, the outcome is within the boundaries set by panel and was reasonable in terms of the performance of the candidates.

The concerns or difficulties raised in the literature about the Hofstee method relate to i) the difficulty judges have when deciding about appropriate failure rates for a test (*cf.* Bandaranayake 2008:840) and ii) when the cumulative frequency score curve misses the rectangle set by the judges and hence, no passing score can be derived (Bowers & Shindoll 1989:5; Norcini 2003:467).

A recent statistical and validity concern about the Hofstee method has been raised regarding its appropriateness and real sensitivity to test difficulty (Tavakol & Dennick 2013:1). Tavakol and Dennick ran multiple simulated test scores through the conventional Hofstee method at varying maximum and minimum pass marks and failure rate and found minimal (1-2%) variation in the eventual pass mark set for the test simulations performed (Tavakol & Dennick 2013:1).

In response to the concerns raised by Tavakol & Dennick, a report was published investigating the use of a modified Hofstee method, which effectively addressed the concerns (Todd, Burr, Whittle & Fairclough 2014:1). Todd *et al.* (2014:1) showed that when setting the maximum and minimum failure rates to 100% and 0%, respectively, (meaning all candidates could pass and fail) and lowering the minimum pass mark to 0% (removing it essentially), the only remaining variable was the upper limit of the pass mark. They then used a fixed percentage under the median (50^{th} percentile) of the group's performance, to set the pass mark. They tested their new modified Hofstee method using the results of more than 50 written assessments in their

institution, with cohorts sizes in excess of 240 candidates, (they specified it needs to be large groups) and reported acceptable pass mark variance of between 5.4 - 8.5%, with a median value of 6.6%.

Todd *et al.* (2014:1) quote the Cohen method as a similar method using percentile benchmarking. However, there is one *significant* difference between their modified Hofstee method and the Cohen method in terms of which percentile is used. The 50th percentile, in a large distribution of selected medical students, is indeed a good and stable marker (similar to the mean) of the *group's* performance, that is, the *whole* group. The Cohen method is not concerned with the whole cohort. It typically uses the 95th percentile to reflect what was possible to achieve in the assessment by the top performing candidates. As such, the 95th percentile is not a whole group performance statistic, it is a reflection of the top performers – which the Cohen method uses as a benchmark of test difficulty (Schoeman 2011:2). The modified Hofstee method described by Todd and colleagues (2014:1) is in essence the same as the Wijnen method, an example of a *pure* norm-referenced method.

The *De Gruijter and Beuk methods* are both similar to the Hofstee method in terms of their basic approach. They also ask a panel of judges to determine pass marks and failure rates for a cohort of candidates sitting a particular test (Beuk 1984; De Gruijter 1985).

However, in *Beuk's* method the panel is asked to provide an expected pass mark and *pass* rate for the given test. The panellists' scores are averaged for their given pass marks and pass rates. The mean values for the pass mark and pass rate are plotted (one mark) on a similar graph as described above in the Hofstee method. Then a diagonal line is drawn through the mark with a gradient determined by standard deviation (SD) of pass rate divided by the SD of pass rate. The point where this line intersects the cumulative frequency score curve of the candidates is set as the pass mark. This method ensures that a pass mark is always derived, since the lines will intersect somewhere on the graph (Bowers & Shindoll 1989:4).

De Gruijter's method is similar to Beuk's methodology, but in this method panellists must provide estimates of a pass mark, *failure* rate (Beuk asks for pass rates) and *certainty* levels for the pass mark and failure rate estimates they provided. All three data points are then used in a complex mathematical equation to set the pass mark,

after incorporating the cumulative frequency score curve of the candidates (Hambleton & Pitoniak 2006:450). The main criticism of De Gruijter's method is the difficulty of explaining the method (poor explicability) to stakeholders (Hambleton & Pitoniak 2006:450).

2.2.8.2 *Relative or norm-referenced methods*

Relative of norm-referenced methods are some of the oldest documented strategies used in medical education to determine who passes or fails (Anonymous author. 1990:444; McManus *et al.* 2005:2; Meskauskas & Webster 1975; Wijnen 1971). In the context of medical education assessments, where the main aim is primarily to *determine competency* in a domain of ability (knowledge, skills and/or attitudes), these methods have limited utility and are deemed not appropriate for use (McKinley & Norcini 2014:97-98). The main reason for this position in the literature is the fact that these methods are most appropriate if *rank ordering and subsequent selection* of candidates is the main purpose of the assessment (Yudkowsky & Downing 2009:L1818).

Translation of the conceptual performance standard into a numerical pass mark score on the observed-score scale of the test (Kane 1998:130), is not explicit when using norm-referenced methods. The *assumption* is made that a certain percentage of candidates are competent and the rest are not and, therefore, fail.

There are two types of norm-referenced methods. They are briefly described here.

Fixed pass rate method

A classic example of this method was the past practice of a fixed pass *rate* of the top 35% of candidates in the MRCP (UK) examinations from 1985 – 2002 (McManus *et al.* 2005:2). The competence of the 36th percentile candidates might have been acceptable, but since they were in a strong cohort, they failed. The reverse circumstance also holds true. Candidates at the 30th percentile may not have been sufficiently knowledgeable, but they would have passed because they were in a weaker cohort of examinees.

Mean-derived method

An example of such a method is the Wijnen method (Wijnen 1971). It uses the overall performance of the group, the *mean* test score, minus two times the SEM to set the pass mark. An alternative to the classic Wijnen method is the mean minus one standard deviation (referred to as the modified Wijnen method in this thesis) (Schuwirth & Van der Vleuten 2010:204). Both of these methods will pass and fail a fixed proportion of the cohort, without a direct reference to competence.

This modified Wijnen method was used for 20 years at Maastricht University medical school for their undergraduate progress test of medical knowledge (Verhoeven, Verwijnen, Muijtjens, Scherpbier & Van der Vleuten 2002:864). This method was also used and described in a research study to compare its outcome to a test-centred method (modified Angoff method) in MCQ tests (George, Haque & Oyebode 2006). The Wijnen method resulted in a 15% higher failure rate compared to the modified-Angoff method (George *et al.* 2006:4).

The classic Wijnen method was used in a study comparing five methods, including the Angoff and Cohen methods to set the pass marks for an OSCE (Kaufman, Mann, Muijtjens & Van der Vleuten 2000). They reported 155 candidates sitting the OSCE, with a mean OSCE performance score of 63.2% (SD=5.3%). The fixed 60% pass mark yielded the highest failure rate (26.5%, n=41) followed by the Wijnen method (8.4%, n=13). The Cohen60 (60% of the 95th percentile) resulted in no failures, while the Angoff method (0.7%, n=1) and the Borderline Group method (1.95%, n=3) had very low failure rates (Kaufman *et al.* 2000:270-271).

The authors reported that the fixed 60% pass mark and norm-referenced Wijnen methods resulted in significantly more failures from the OSCE. The other three methods (Angoff, Cohen60 and Borderline Group) had similar results (very low failure rates). The Angoff and Borderline Group methods were considered reasonable and defensible approaches. It was interesting that the authors did not include the Cohen60 method in their conclusion about the reasonable and defensible methods, although all *three* resulted in very similar failures (n=0, 1, 3 respectively). The narrow distribution of OSCE scores (SD=5.34%) was reported as the most likely reason why the Cohen60 method had no failures. The authors did call for more research on the Cohen60 method, but from the results of this study it seems that the Cohen60 method

performed similarly to the other two methods in terms of outcomes (Kaufman *et al.* 2000:270-271).

Although relative methods have been criticised for not being suitable for competency assessments, they do however, have one positive feature. They are directly sensitive to the *difficulty* of the test as perceived by the *cohort under assessment.* It is this specific attribute of norm-referenced methods that compromise methods try to harness and incorporate into the process of setting pass standards.

2.2.9 The Angoff method of standard setting

The Angoff method (Angoff 1971:514-515) probably is the best known and most widely used absolute PBTC standard setting method in medical education internationally (Cusimano 1996:113; Downing *et al.* 2006:52-53; Norcini 2003:465). In the subsequent sub-sections, various aspects are described and discussed to provide an overview of the method. The first sub-section introduces the Angoff method and discusses its historical origin and basic methodology. Thereafter, some of the common modifications of the Angoff method are described and explained, ending with the Yes/No version of the Angoff method. Finally, its use and application in different assessment formats are described, including some research studies that compared the Angoff method's performance to other standard setting methods. The comparisons are outlined in a focused, tabulated format.

2.2.9.1 *Introduction and basic principles of the Angoff method*

The Angoff method was originally described by William Angoff in 1971 (Angoff 1971:515). He described two basic versions of his method, one in the main text of his paper and the other in a footnote at the bottom of the page. Interestingly, it was the 'footnote' that became the commonly known 'Angoff method' (Clauser *et al.* 2009b:390) and the one described in the main text, the *original* version, became known as the 'Yes/No' Angoff method (Downing *et al.* 2006:53).

The *basic methodology* of both versions of Angoff's method is the same (Cizek *et al.* 2004:40; Clauser, Clauser & Hambleton 2014:20). It is regarded as the prototype absolute/criterion-referenced standard setting method (Yudkowsky & Downing 2009:L2003). It is based on the classic approach using a panel-based, item-by-item

reviewed, test-centred process capturing the collective judgements of a panel of experts on the predicted performance of a hypothetical borderline candidate/or set of candidates on each test item (Brandon 2004:60; Clauser *et al.* 2009b:390).

The procedure to apply the Angoff method starts with the appropriate selection and training of the panellists/judges (Hambleton & Pitoniak 2006:436). An important aspect in the training of the judges is having a discussion and coming to a shared understanding of what the characteristics and abilities of a hypothetical 'borderline candidate(s)' in the assessed cohort would be. It is this shared or agreed understanding of characteristics and abilities of a borderline candidate, that will be used by the panel to judge the items (Hambleton & Pitoniak 2006:437). The specific practical procedural steps should also be explained and clarified, such as the purpose and nature of the test, how judgements will be gathered and the specific judgemental task given to the panellists (Hambleton & Pitoniak 2006:437; Kane 1994:441). Most of the common modifications of the Angoff method occur during these procedural steps, in an effort to increase panellists' agreement and realism on item difficulty judgements (Ricker 2006:55). These modifications are discussed in the next sub-section.

The exact phrasing of the task or instruction given to the panellists varies, however, the basic panel instruction in the Angoff method can be summarised as: "Please use your judgement to answer the following question: What proportion (or percentage) of a 100 hypothetical 'borderline' (just-competent) candidates *would* answer each test item correctly" (Kane 1994:429; Ricker 2006:54). The panellists' scores are then collated and averaged for each item to determine the mean Angoff rating or judgement *per item*. Subsequently, the individual mean item ratings are averaged across all the items to determine the pass mark for the paper (Downing *et al.* 2006:53). Alternatively, the mean Angoff judgement of the test, across all the test items, is calculated *per judge* and the mean of all the judges' final scores becomes the Angoff pass mark of the test (Ricker 2006:55).

Panel composition and size

The critical role and functioning of the *panel of judges* in the Angoff method, and in any other test-centred standard setting method, has been widely established and referred to in this thesis. The composition and size of the judging panel are therefore two important aspects, which contribute to the Angoff method's interval validity process and evidence (Verheggen, Muijtjens, Van Os & Schuwirth 2008:204). The use

of *content experts* on judging panels has long been advocated in the credentialing and licensing standard setting literature (Hambleton & Pitoniak 2006:436; Norcini & Shea 1997:41). This recommendation was reviewed by Brandon in his seminal literature review of the modified Angoff methods and he reported, that content expertise does seem to affect item difficulty estimations (Brandon 2004:66). Generally, the greater the content expertise of panellists, the easier the test items are perceived and rated, resulting in unrealistically high pass marks (Livingston & Zieky 1989:137; Norcini *et al.* 1988:63; Verheggen *et al.* 2008:209; Verhoeven *et al.* 2002:866). Brandon advised that not all panellists needed to have high levels of expertise and that judges who have a general knowledge of the test content (generalists) would be equally appropriate, since the effect of iterative discussion rounds with normative data reduced the differences between content experts and generalists (Brandon 2004:66).

Additional points raised by the literature that were also important for panellists were a good understanding of the examinees and the educational context of the test (Verheggen *et al.* 2008:205).

Verheggen *et al.* (2008:209) raised an important aspect in their study. They studied the effect of judges' item *knowledge* on their subsequent Angoff ratings and stringency as judges. They asked a panel of 13 judges (postgraduate psychiatry trainees), who were familiar with the content area of the items (undergraduate psychiatry knowledge), to provide Angoff item difficulty estimates, but also to answer the test items, similar to the candidates, in addition. The authors found that item content knowledge levels had a considerable effect on their subsequent Angoff ratings and stringency.

Following these findings, Verheggen and colleagues (2008:209) raised the question of should Angoff judges be provided with *item answer keys* during rating rounds? If judges *were not* given answer keys and they don't know the answer to some test items themselves, they seem to rate them as more difficult than those judges who do know the answer. This leads to increased inter-rater variability in the ratings and, therefore, reduced reliability of the Angoff method, with less generalisability of the resultant Angoff pass mark, which is undesirable (Verheggen *et al.* 2008:209). If the Angoff judges *did* receive the item answer keys, the authors argue, in line with the rest of the literature, that the items will seem easier to the judges and hence item difficulties will be judged lower than what the actual case might be and lead to unrealistically high

pass marks and resultant failure rates (Verheggen *et al.* 2008:209). They conclude that providing judges with item answer keys, would enhance the reliability of the Angoff ratings, but probably at the expense of making them less valid, since "the judgements would be consistently off the mark" (Verheggen *et al.* 2008:210).

The authors suggest a complex alternative strategy of not providing the answer keys to judges during the Angoff process, making the judges also answer the items of the test (like the candidates) and then using regression analysis, similar to what they described in the paper, for those test items the judges gave incorrect answers to calculate the eventual Angoff pass mark (Verheggen *et al.* 2008:210). This strategy seems unlikely to attract willing judges, who would consent to sit the test first, since they might fear the possible humiliation of scoring poorly on the test themselves.

Another study investigating the effects of providing answer keys during the Angoff process, reported that judges who received the answer keys rated hard items more difficult and easy items more easy, than judges who did not have the answer keys (Hudson (Jnr.) & Campion 1994:863). Judges with answer keys therefore had a wider range of item difficulty ratings compared to judges without the answer keys. The authors concluded that provision of answer keys can influence item difficulty ratings in Angoff processes. If the test, with answer keys, appears easier to the judges, they set higher pass marks for the test. These findings by Hudson & Campion (1994:863) therefore compliment the findings as reported by Verheggen et al. (2008:210).

The optimal *size of the Angoff panel* has been an ongoing debate in the literature, with the Brandon reporting studies advising from 5 to 30 panellists (*cf.* Brandon 2004:67). Downing et al. (2006:51) advised that from a practical feasibility perspective 5-6 judges are a minimum and 10-12 is the maximum required for a credible outcome. However, after a systematic review of the literature, Brandon recommended there should be at least 10, with 15-20 being the ideal number (Brandon 2004:68).

2.2.9.2 *Modifications of the Angoff method*

Given the vast amount of published research papers, reviews and comparisons to other methods over the past 40 years, justifies the conclusion that educators hold the Angoff method in high regard (Brandon 2004:60; Clauser *et al.* 2009b:390; Kane 1994:439).

However, the Angoff method is not without problems or concerns. *Multiple* modifications to the basic Angoff method over time are related to the *limited* ability of judges and judging panels to accurately predict the performance of hypothetical borderline candidates on test items (Chinn & Hertz 2002:3; Mee, Clauser & Margolis 2013:27). Hambleton and Pitoniak (2006:440) suggested that there might be a 100 or more 'modified Angoff methods' used in standard setting practices. The *commonest* modifications to the original Angoff method are only briefly discussed here, since many have already been mentioned in preceding sections describing the challenges of PBTC methods in general. The central aim of modifications has been to increase the accuracy (validity) and consistency or agreement (reliability) of panel ratings so as to produce a more defensible pass mark (Mee *et al.* 2013). These modifications are all part, to some extent or another, to references in the literature to the *modified* Angoff method (Brandon 2004:60; Ricker 2006:55).

Common modifications

The use of *multiple iterative discussion rounds* during the Angoff standard setting process, with review of ratings and subsequent re-rating, is a common occurrence (Ricker 2006:55). This approach seeks to improve agreement and build greater consensus amongst the panellists regarding item difficulty estimates. Although this strategy appears to be a good way of improving the validity and reliability of the judges' ratings, Mee *et al.* (2013:34) and other authors (Clauser, Harik, Margolis, McManus, Mollon, Chis & Williams 2009a) reported that studies done at the NBME on Angoff standard setting of medical licensing examinations, showed that while iterative discussions and consensus building, in the *absence* of actual performance data, does lead to greater convergence of test item estimates, it may exacerbate inaccuracies with regard to test item difficulties. So, while the opinions of panellists converge, they become less accurate at predicting the performance of borderline examinees.

Hurtz and Auerbach (2003) reported a similar finding in their meta-analysis study on the effects of some modifications of the Angoff method. They found that discussions amongst the panellists about their item ratings, with subsequent re-ratings, generally resulted in *higher* pass marks being set, (Hurtz & Auerbach 2003:596) and reported uncertainty whether these higher resultant pass marks were more or less valid. They did report their particular concerns about the influence of 'group member personalities' on the eventual pass mark and how hard it would be to control these personality influences and effects (Hurtz & Auerbach 2003:597).

This modification of the Angoff method is labour intensive and time-consuming (Cusimano 1996:S116; Norcini & Shea 1997:44) and therefore, might not be feasible to implement in some contexts. The alternative is to ensure adequate numbers of trained judges on the Angoff panel, as mentioned in the preceding section. Their individual ratings are simply averaged across the panel to deal with divergent ratings on individual items (Berk 1996:221; Norcini 2003:465).

Another approach to modifying the Angoff method has been the use of *normative performance data* of the test items during the discussion rounds, in an attempt to temper the unrealistic expectations of panellists (Ricker 2006:55). It is commonly referred to as the 'reality check' (Livingston & Zieky 1982:57). Without access to normative data during the discussion rounds, most panels produce item difficulty estimates that bear little or modest resemblance to the actual empirical item difficulty data (Clauser *et al.* 2009a). When normative data are provided during the discussion rounds, panellists change their judgements to improve alignment with the provided data, usually in a downwards direction (Hurtz & Auerbach 2003:597). This defeats the purpose of the Angoff procedure and the process essentially becomes a compromise method (due to the norm-referenced input) (Mee *et al.* 2013:33).

In a review of 11 studies of modified Angoff methods, Brandon (2004:77) reported that the mean improvement in the correlation coefficients between the judges' estimates and actual item difficulty was 0.20 (SD=0.14), after review of the normative performance data. This finding is supported by another study, who reported that the use of normative data in the standard setting process, helped judges to become more accurate in their item difficulty estimates for borderline students (Cusimano & Rothman 2003:S89; *cf.* Verheggen *et al.* 2008:210).

Kane also made the point that judges usually set unrealistically high pass standards, if they are not provided with some 'impact data' (data which informs them of the failure rate consequences of their intended pass standard) during the process, and have a chance to review it (Kane 1998:136).

The Yes/No Angoff method

The final common modification which is discussed relates to the Yes/No version of the Angoff method, which was briefly mentioned in the introductory paragraph on the Angoff method in section 2.2.9.1 above. As explained before, the Yes/No Angoff

method is actually regarded as William Angoff's *original* description of his standard setting method, however it is lesser known and used than the 'footnote' or classic Angoff method.

Downing *et al*. (2006:53) explains the procedure of applying Angoff's *original* method (Angoff 1971:514-515) of setting the absolute pass mark of a written paper as follows: "A variant of the Angoff method (actually Angoff's original method) is to ask judges to make a simple 'yes' or 'no' judgment about each item/prompt/question. The question becomes, 'Will the borderline examinee respond correctly to this item?' All 'yes' answers are coded as 1, with 'no' answers coded 0. The simple sum of the 1s and 0s becomes the raw passing score when averaged over all judges. This simplified Angoff method (direct or Yes/No method) may be useful for some types of examinations, such as laboratory tests, for which use of the traditional Angoff method would be difficult."

The appealing feature of the Yes/No Angoff method is its simplified judgement task to panellists (Cizek *et al.* 2004:42). Research comparing the traditional Angoff method and the Yes/No version reported that panellists found the Yes/No version easier to use, less sensitive to normative data and had lower intra-panellist variation and concluded, that the Yes/No version may produce more valid pass marks than the traditional Angoff method. (Impara & Plake 1997:363).

The Yes/No method can be applied in two ways, in terms of the directions given to panellists about the visualised target they must use when making predictions. One is the discussed and agreed *hypothetical* borderline candidate in the target group or alternatively, the panellists may be directed to think of an *actual* borderline candidate familiar to each of them (Impara & Plake 1998:355). The classic 'group of 100 *hypothetical* borderline candidates' strategy is, therefore, not routinely used with the Yes/No method (Cizek *et al.* 2004:42). This simplified judging task used in the Yes/No Angoff method, as explained above, is especially useful and appealing when novice judges, who are new to standard setting, are used to set the pass standard.

Impara and Plake (1998:363) concluded from their study that when using the Yes/No Angoff method, using only one round of judgements may be permissible and that normative data is less influential and important, compared to the traditional/classic Angoff method. The Yes/No method was supported for implementation in comparison to other methods in previous studies (Downing, Lieska & Raible 2003:S87). The

Yes/No Angoff method is also the method the American Board of Internal Medicine (ABIM) uses to derive their passing scores (Tormey 2014:8).

Use and implementation of the Angoff method

The Angoff method was originally designed for selected-response (MCQ-type) written tests (Angoff 1971:515; Clauser *et al.* 2009b:390), but has since been used extensively with constructed-response written tests (e.g. short answer questions) (Cizek *et al.* 2004:40) and in performance-based assessment, such as OSCEs (Boursicot, Roberts & Pell 2006; Kane 1998:141). The modifications mentioned above supported its increased flexibility and application in multiple testing contexts.

As mentioned previously in this chapter, the literature is clear about the fact that different standard setting methods result in different passing scores. However, comparing the outcomes of different standard setting methods with each other is also regarded as an important source of external validity evidence for standard setting methods' outcomes (Kane 1994:449).

Given that the Angoff method and its modifications are the most researched method of all standard setting methods (Brandon 2004:80), providing an exhaustive list and discussion on all the comparisons to the Angoff method, falls outside the remit of this thesis. However, since the present study and thesis did compare the Cohen and Angoff methods, a tabulated description of comparative studies involving the Angoff method is provided in Table 2.1. The only other published study comparing the outcomes of the Angoff and Cohen methods on written tests in medical education is mentioned below, but discussed in the next section under the Cohen method.

TABLE 2.1: PUBLISHED ANGOFF METHOD COMPARISONS AND REVIEWS

Written assessment context	
Compared to	**Source**
Angoff (Yes/No)	Chinn & Hertz 2002; Impara & Plake 1997; 1998
Nedelsky	Chang 1999
Bookmark	Buckendahl *et al.* 2002; Cetin & Gelbal 2013; MacCann & Stanley 2010; Peterson *et al.* 2011; Smith, Davis-Becker & O'Leary 2014
Bookmark (vs. Yes/No Angoff)	Hsieh 2013
Hofstee/Beuk	Bowers & Shindoll 1989
Hofstee	Parmar, Shah & Parmar 2014; Stern, Friedman Ben-David, De Champlain, Hodges, Wojtczak & Schwarz 2005
Nedelsky/Borderline group/Contrasting groups	Livingston & Zieky 1989
Nedelsky/Hofstee/Ebel	Downing *et al.* 2003
Wijnen	George *et al.* 2006
Cluster analysis	Hess, Subhiyah & Giordano 2007
Performance-based context	
Compared to	**Source**
Ebel	Cusimano & Rothman 2003
Ebel (vs. Yes/No Angoff)	Yudkowsky & Downing 2008
Hofstee	Wayne *et al.* 2005
Borderline Regression	Hobma, Ram, Muijtjens, Grol & Van Der Vleuten 2004
Review papers dedicated to Angoff	
Time frame/Focus	**Source**
1971 to 2000	Hurtz & Auerbach 2003
1971 to 2004	Brandon 2004
Modifications	Katz & Tannenbaum 2014; Norcini, Lipner, Langdon & Strecker 1987; Ricker 2006

2.2.10 The Cohen method of standard setting

The Cohen method of standard setting was first described by Janke Cohen-Schotanus and her colleagues in 1996 in the Dutch literature (Cohen-Schotanus *et al.* 1996). It was subsequently published in the English literature 14 years later (Cohen-Schotanus & Van der Vleuten 2010).

In the 2010 publication of the Cohen method, the authors explain the reason for developing the method. It was due to the problems they faced with their existing standard setting methods at their home institutions (Cohen-Schotanus & Van der Vleuten 2010:157).

Maastricht University used the modified Wijnen method, which was well described previously in this chapter (section 2.2.8.2), leading to stable failure rates of 17%, but considerable fluctuations in the resultant pass mark, ranging from 15-46% in 54 local MCQ knowledge tests for first year medical students (Cohen-Schotanus & Van der Vleuten 2010:155). The opposite experience was the case at Groningen University where they used a fixed 60% pass mark, after correction for guessing. On applying this standard to 52 local MCQ knowledge tests for third and fourth year medical students, it resulted in a failure rate ranging from 17-97%, with a mean of 53%. However, year after year, the two institutions' final year students performed very similarly in the national progress test, which revealed that the students' mean knowledge base were of a similar standard (Cohen-Schotanus & Van der Vleuten 2010:156).

The authors realised that both institutions needed a different standard setting method. Norm-referencing using the mean as a reference point was viewed as unacceptable due to the influence that under prepared students would have on it and because a fixed number of candidates would fail every time, which was seen as unfair (Cohen-Schotanus & Van der Vleuten 2010:157). The fixed 60%, after correction for guessing was viewed as unacceptable, since it was insensitive to the difficulty of the test and hence indefensible (Cohen-Schotanus & Van der Vleuten 2010:157).

A compromise method was needed that took account of the difficulty of the test, but which would set an absolute pass mark, where all students who scored above the criterion could pass the test. The authors realised that the top performing candidates

provided the answer and was "one stable factor in the complicated process of standard setting" (Cohen-Schotanus & Van der Vleuten 2010:157). Educators and students found the Cohen method's methodology acceptable and using the 95th percentile as a benchmark of test difficulty, with 60% of the benchmark as the pass mark (Cohen60 model), after correction for guessing, resulted in acceptable pass marks and failure rates at Groningen (Cohen-Schotanus & Van der Vleuten 2010:159). The authors also commented on the Cohen method's favourable cost- and time effectiveness (Cohen-Schotanus & Van der Vleuten 2010:159).

Since the original publication of the method, the Cohen method has been compared, in terms of performance, to other standard setting methods. Three studies were identified in the literature. The study by Kaufman et al. (2000) has already been reviewed earlier in this chapter (section 2.2.8.2) under the mean-derived norm-referenced method, since the Wijnen method was part of that study. The other two studies involving the Cohen method was in the context of first year undergraduate law students in Belgium (Dochy, Kyndt, Baeten, Pottier & Veestraeten 2009) and first and second year undergraduate medical students in the UK (Taylor 2011).

In the paper by Dochy et al. (2009), the authors studied nine standard setting methods, which they classified into three groups as Absolute (6 methods), Relative (2 methods) and Mixed (1 method). The Cohen method was the only method included in the "Mixed" group. They also used the Cohen60 model, similar to the original Dutch paper (Cohen-Schotanus & Van der Vleuten 2010). The study had a complex design and methodology with the aim to compare the effect of the nine methods on the size and composition of the borderline group and the discrimination between different types of students in terms of pass and fail classifications (Dochy *et al.* 2009:176-177). The authors reported that in their study, based on the two written tests (mixture of MCQ and Short Answer Questions) they used, they found it hard to identify one method that was the best discriminator between the students (Dochy *et al.* 2009:181). The method that performed the best was a complex test-centred method, referred to as the Method of Cascallar and Cascallar (*cf.* Dochy *et al.* 2009:175). Two other methods, including the Cohen60, were also reported as performing "reasonably well" (Dochy *et al.* 2009:181).

The third study by Taylor (2011) was the paper that aligned best to the present study. It investigated the Cohen method's methodology and rationale by using a modified

Angoff method (which included panel discussions) as a comparative benchmark. In essence, Taylor modelled the Cohen method to produce a similar outcome as the Angoff method. The study questioned the use of correction for guessing on the MCQs and the subjectivity of the 60% multiplier in the Cohen method formula, which Cohen-Schotanus & Van der Vleuten used in their 2010 paper (Taylor 2011:e678):

$$\text{Pass mark (\%)} = C\% + 0.6(P-C)$$

Where C= expected % from random guessing (defined as 1/number of MCQ options, converted to %) and P= the 95th percentile point (%).

Taylor seems to have made two assumption errors. First, the methodology used by the Dutch to correct for guessing is unique in the sense that it does not add CIV to the test scores. It actually raises the pass mark to account for guessing, but does not penalise the students for incorrect answers on the MCQ scores. The literature on formula scoring or negative marking, as presented and discussed in this chapter, is critical of this practice due to the adding of CIV when candidates' scores are altered due to negative marking, which is not the case here. Second, the 0.6 or 60% multiplier is not chosen with blind subjectivity as suggested by Taylor, it is actually the Dutch policy on expected minimum knowledge on MCQ tests (Cohen-Schotanus & Van der Vleuten 2010:157), which Cohen-Schotanus & Van der Vleuten incorporated into their Cohen-method formula to still use the national policy, but in a fairer way (by also correcting for the difficulty of the test).

Taylor derived a modified (simplified) Cohen method formula, after removing correction for guessing. The new formula was:

$$\text{Pass mark} = K \times P.$$

Where K= the multiplier percentage (%) and P= the percentile used as difficulty benchmark. She calculated the optimum K and P in their context, based on local tests, which had been subjected to a modified Angoff method and cumulative density functions (CDF) of the test data. Taylor reported that 65% of the 90th percentile gave the closest outcome to the Angoff method, which was regarded as the criterion-referenced standard (Taylor 2011:e680).

They also used tracker MCQ items to evaluate if the 90th percentile was stable over the course of different students sitting the same tests. The cohort sizes in this study were large (n=370) and Taylor did comment on the fact that more research is needed to test the Cohen method on smaller cohorts and determine if it affects the validity of the method (Taylor 2011:e681). The present study also used tracker MCQ items and had smaller cohorts, albeit postgraduate cohorts, which should address this research need expressed by Taylor.

Taylor's study provided an alternative way to derive the multiplier % in the Cohen method – by benchmarking it from an alternative standard setting method. The use of a policy criterion, as done by the Cohen-Schotanus & Van der Vleuten in the 2010 original Cohen method paper is equally justified, as long as there is correction for difficulty incorporated in the method.

Appeal of the Cohen method

After a review of the literature on the Cohen method, certain aspects have emerged that make this method appealing as a standard setting alternative. It had good feasibility and sustainability credentials, especially in resource-constrained contexts like South Africa. It is also sensitive to test difficulty and uses the top performing candidates to benchmark the difficulty of the test, which is unaffected by the *mean* cohort score. The Cohen method sets an *absolute* pass mark, (Dochy *et al.* 2009:176), which all students can achieve (Kaufman *et al.* 2000:270). The faculty's expected *minimum* standard of content mastery can be expressed in the method. For example, in the case of Groningen, students must achieve 60% MCQ items correct, after correction for guessing and difficulty (Cohen-Schotanus & Van der Vleuten 2010:157). It can alternatively be modelled to reflect the outcomes of other standard setting methods, such as the Angoff method in the case of the Taylor paper (Taylor 2011). Retrospective data modelling on previous test results is possible to derive at optimum model to use, which emphasizes the flexibility of the method.

Challenges identified in relation to the Cohen method

Although the Cohen method has strong potential as a useful compromise method, some challenges have been identified in the literature review that need acknowledgement. The Cohen method still has a small research base behind it and needs more research exposure in different contexts, such as postgraduate medical education, with smaller cohorts and in resource-constrained settings. The stability of

top performing candidates, in different contexts, is largely unknown, although Taylor's paper did provide some evidence that they are stable in large cohorts of undergraduate medical students (Taylor 2011:e680). The minimum cohort size to enable the appropriate and valid use of the method needs clarification. The Cohen method is potentially vulnerable to manipulation of the pass mark by the top candidates (if they collude). Although this is extremely unlikely to occur, it is theoretically possible, however the risk is similar to other standard setting methods where candidate data is used to derive or influence the pass mark, e.g. modified Angoff (with normative data), Wijnen method or Hofstee method.

In summary, the Cohen method is a compromise method that encapsulates the advantage of norm-referenced methods' ability to incorporate the relative difficulty of the assessment and therefore, can correct or compensate for educational errors that could have entered into the system and have negatively affected the performance of the candidates sitting the assessment. It also benefits from the advantage and ability of criterion-referenced methods to set an absolute pass standard, where all students can potentially pass the assessment. There is no fixed failure rate. The method can be used either in conjunction with correction for guessing, in the case of MCQ tests, or without it. The method's application and use in different contexts needs further research to test its assumptions.

2.2.11 The social accountability of standard setting in medical education

The concept of the 'social accountability of medical schools' has been an ever increasing force in the literature and in global medical education and healthcare meetings for the last 20 years (Gibbs 2011:605; Woollard & Boelen 2012:22). Boelen and Heck (1995), on behalf of the World Health Organisation (WHO), published their seminal paper on defining social accountability (SA) for medical schools and how they (the medical schools) can measure their efforts and progress to advance health outcomes in their contexts. This paper made a strong case primarily to medical schools, but also to their regulators and governments, to recognise their significant influence and role in responding to the priority health needs of their regions and nations, and therefore need to accept their "accountability" to societies they are mandated to serve.

Social accountability for medical schools was *defined* by the WHO as "'the obligation to direct their education, research and service activities towards addressing the priority health concerns of the community, region, or nation they have a mandate to serve. The priority health concerns are to be identified jointly by governments, health care organisations, health professionals and the public" (Boelen & Heck 1995:3).

Medical schools, as well as all other health education and training institutions, are therefore called upon to directly respond to the priority health needs of societies and not merely be beacons of academic independence and autonomy, but, together with all other health stakeholders reform their educational endeavours, research foci and service delivery efforts to actively promote health outcomes for all (Boelen & Woollard 2009:888).

The *values of SA* were derived from the same values as the WHO expressed towards the vision and goal of global 'health for all', which were endorsed by all member states and nations in 1981 (Boelen & Heck 1995:1). They are *relevance, quality, equity and cost-effectiveness.* These values became the global moral and ethical driving forces to reform the activities of medical schools towards becoming more socially accountable to the societies they serve.

One of the key indicators of whether a country has the capacity to meet its healthcare needs and goals, is the number *and quality* of key healthcare workers (HCWs), and especially the number of doctors, nurses and midwifes working in the country (WHO 2006:xv). The WHO calls it "the human link that connects knowledge to health action" (WHO 2006:xv). Figure 2.7 demonstrates the relationship of density of HCWs to the probability of survival.

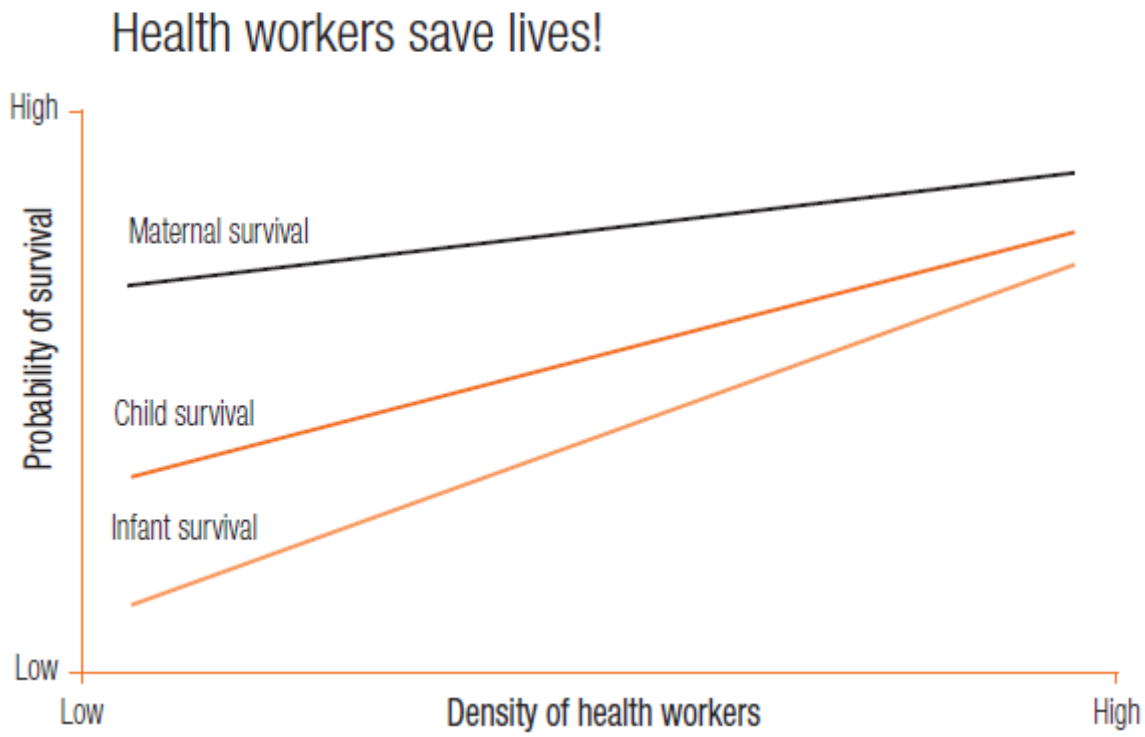


FIGURE 2.7: THE IMPACT OF THE NUMBER AND QUALITY OF HEALTH WORKERS ON HEALTH OUTCOMES
[Taken from the World health report 2006 (WHO 2006:xvi)]

In their 2006 World health report, the WHO identified 57 countries around the world with a critical shortage of a total of 2.4 million doctors, nurses and midwifes, the vast majority of these located in sub-Saharan Africa (36 countries) (WHO 2006:xviii). The global shortage of all HCW was estimated to be about 4.3 million (WHO 2006:15).

A global SA response and action was called for by Gibbs and McLean to meet the shortages of HCW and to "level the playing fields" (Gibbs & McLean 2011). They discussed a multitude of factors contributing to the global challenges and inequalities in HCW distribution and shortages, from the 'brain drain' from poorer to richer countries, wars and corruption to international sanctions with lack of opportunities (Gibbs & McLean 2011:621). The call to increase global production of HCWs to meet the demand was also echoed by other authors (Norcini & Banda 2011:83).

In a similar response to authors in the medical education community, the WHO published a report in 2013 entitled: 'Transforming and scaling up health professionals' education and training guidelines' (WHO 2013), which is an earnest call and comprehensive guide to member states to train and produce more HCWs to meet the global demand for HCW and especially to reduce the critical shortage of HCWs in the

57 countries named in the 2006 World health report. The importance of both quantity *and quality* of HCW were emphasised in the report (WHO 2013:24). Accreditation and regulation was specifically mentioned as a driver of quality and a responsibility of the state.

The state should therefore facilitate quality education of its health professionals as well as ensure that sufficient HCWs are being trained to meet the needs of the country (WHO 2013:24). In fact, one of the key policy issues from the document was the protection of the public from incompetent providers, by ensuring minimum qualification standards were maintained by educational institutions (WHO 2013:24). It is this interface between *the quantity and the quality* of the HCWs trained and produced in a country, where educational institutions should aim to get the balance right – producing maximum numbers, but ensuring robust quality assurance mechanisms are in place to protect the public and maintain trust in the health professions (Norcini & Banda 2011:83; WHO 2013:26). This is an important focus of the social accountability obligation of health professions' training institutions within a country and needs to be executed in an equitable, relevant and cost-effective manner (Boelen & Woollard 2011:615). Boelen and Woollard (2011:615) referred to this envisioned system as finding the "silver bullet" for medical education institutions, when they produce "the right doctors to practice the right medicine with the right partners at the right time in the right place".

It is this "producing and developing the goods" for a country and the world, which encapsulates precisely the potential and ability that a socially accountable medical education system has, according to Gibbs (2011:605). He goes on in his commentary in a special issue of the *Medical Teacher* journal dedicated to SA, that all organisations involved in the development of healthcare professionals share a *global* responsibility to be socially accountable, both locally and internationally (Gibbs 2011:606).

The Global Consensus on Social Accountability of Medical Schools (GCSA) agreement document was published in 2010, to consolidate a two-year, worldwide Delphi consensus building process involving 130 organisations (GCSA 2010:1). The final part of the consensus building process was concluded with a conference in East London, Eastern Cape Province, South Africa in October 2010. Ironically, the conference was held in conjunction with the 25th anniversary of the Walter Sisulu University School of Medicine (GCSA 2010:15), a medical school founded in 1985 against all odds under the

rule of the Apartheid government of South Africa. It had a specific social accountability mandate to address the health needs of the nearly 4 million, predominantly black inhabitants, of the rural Transkei homeland (which is now part of the Eastern Cape province of South Africa) (Kwizera & Iputo 2011:651). Up to 1994, South Africa was ruled by the infamous Apartheid regime and thus by its mere nature could not be socially accountable to *all* its citizens (Kwizera & Iputo 2011:651). However, 16 years after the establishment of democratic South Africa in 1994, the world signed a global consensus document on social accountability in the very country which epitomised the opposite policy for decades, a truly remarkable occasion.

The GCSA agreement delineated ten areas, or directions, for medical schools and accreditation bodies to engage with to develop and measure their SA obligation. Two of the areas (Areas 6 and 7) addressed, *inter alia,* the quality assurance of standards in the education of doctors (GCSA 2010:8-9).

Social accountability and standard setting in the CMSA

As demonstrated above, the quality assurance of medical graduates as well as producing adequate numbers of well-trained doctors, at all levels, is an important social accountability application of medical schools. This aspect of SA is emphasized by Woollard and Boelen (2012:24) when they explain the "quality cascade" model, as used in Indonesia's accreditation system. Here quality assurance on every level of the medical education and training spectrum is a focus point with the end goal of the cascade being quality practice by the graduates, leading to the best health outcomes. Although it is hard to prove the direct relationship between these quality assurance processes and the eventual improved health outcomes, SA mandates the development of these processes since they are *most likely* to contribute to improved health for all in the long run (Woollard & Boelen 2012:24).

Therefore, the appropriate medical education curricula, processes and assessment strategies of graduates are very much aligned with the ethos and values of SA (Lindgren & Karle 2011:668). This aspect of SA has been recognised and incorporated by the World Federation for Medical Education (WFME), whose aim is to provide international standards for medical education and accreditation of medical schools (Boelen, Dharamsi & Gibbs 2012:181; Lindgren *et al.* 2012:11; Lindgren & Karle 2011:668-671).

As explained in Chapter 1, the CMSA is now the mandated national postgraduate medical and dental licensing examinations body in South Africa (HPCSA & CMSA 2014). The CMSA differs from the medical schools in South Africa, since it does not have a training role in the postgraduate medical and dental education process, but functions purely as an assessment organisation. As a result of this focused function and mandate in the educational process of licensing specialist doctors in South Africa, its SA translation is also unique.

The Conceptualisation–Production–Usability (CPU) model is proposed by Boelen & Woollard (2009:890) as a framework of three domains guiding institutions to deliver their SA obligation to the societies they are mandated to serve.

Boelen & Woollard (2009:890) describe the domains as follows: "The domain of conceptualisation involves the collaborative design of the kind of professional needed and the system that will utilise his or her skills. The domain of production involves the main components of training and learning. The domain of usability involves initiatives taken by the institution to ensure that its trained professionals are put to their highest and best use."

From the preceding text, it is clear that the CMSA (and CoP) has no role in the Conceptualisation or Usability domains, but has a major role in the Production domain of specialist doctors in South Africa.

Social responsibility of the CMSA

The social responsibility engagement of an institution is the first level of engaging with its SA obligation (Boelen & Woollard 2011:616). This is defined as recognising and being aware of your duties towards being a socially accountable institution (Boelen *et al.* 2012:181). In the context of the CMSA and, more specifically to the scope of this thesis, the CoP, this would mean that the CMSA and CoP are aware that they are the gatekeepers of the quantity and the quality of graduates who are licensed to practice as specialists (production domain). Their examinations are all high-stakes and have major implications for the passing or failing candidate, as well as on the public who need the health services of the specialists and will be exposed to their skills.

Social responsiveness of the CMSA

The next level of SA engagement is that of social responsiveness, which is defined as committing to a course of action to respond to the social responsibility of an institution (Boelen & Woollard 2011:615). For the CMSA and CoP, this would translate to responding to their gatekeeper responsibility by setting and administering high quality examinations and subsequently setting appropriate and defensible passing standards for their examinations.

This is the SA level to which standard setting and this thesis speaks. The next step is to translate the SA values (quality, relevance, equity and cost-effectiveness), which inform this 'Educational production domain' to the standard setting sphere for the CoP and CMSA. This would answer the question of: "How does socially accountable standard setting look in the CoP?" Table 2.2 presents the translation of SA values to standard setting.

Table 2.2 shows that there is a strong link between SA values and the three utility pillars of a standard setting strategy – Robust, Responsible and Realistic. If the CMSA and CoP work towards implementing these three utility pillars of standard setting and its underpinning parameters, then they will become increasingly socially accountable in their assessment mandate to the South African society.

The leadership and management of the CMSA and CoP are of utmost importance and ultimately responsible for the long-term quality development of assessment and standard setting systems in the organisation (Lindgren & Karle 2011:668; Norcini & Banda 2011:83). As Lindgren & Karle (2011:668) state: "Quality is built from the bottom, but must be supported from above, with a clear long-term vision for reform and development".

TABLE 2.2: TRANSLATION OF SOCIAL ACCOUNTABILITY VALUES TO STANDARD SETTING

SA value	Healthcare Definition (Boelen & Woollard 2011:615)	Standard setting translation
Quality	Healthcare is person-centred and interventions are coordinated and relevant to serve the needs of the patient.	The quality of the standard setting method's results are directly related to its defensibility as a standard setting method, which in turn rests on its credibility (validity and reliability) and acceptability of failure rates and objectivity. The quality of the Standard setting method boils down to how **ROBUST** the method is in the local context
Relevance	The most prevalent and important priority health issues are addressed first and foremost.	The priorities of assessment strategies need to be set in a **RESPONSIBLE** manner. High quality assessment processes, instruments and items are essential before a standard setting strategy can be used to determine pass marks. It is not appropriate to attempt a standard setting procedure using poor quality tests items/processes that have questionable credibility. The eventual pass/fail decision will be jeopardised by the weak assessment strategy used.
Equity	Every person in society has access to and can benefit from essential health services.	This aspect relates to fairness of the standard setting method to candidates and patients. How well can stakeholders understand how the method works (explicability) and how transparent is its methodology. Does the method behave in a **RESPONSIBLE** manner in the local context?
Cost-effectiveness	Available resources are best used to achieve maximum health benefits to both the individual as well as the population	The cost-effectiveness of standard setting method relates to its practicability (feasibility and sustainability), which is a reflection of how **REALISTIC** is its use in the local context. The implemented method needs to provide the pass standard (Norcini & Banda 2011:84)

2.2.12 The utility of a standard setting method

Building on the arguments raised in this chapter thus far, there are essentially two ways of determining the difficulty of a test or assessment task: i) ask a panel of experts in the field to evaluate the test content, mostly item by item (test-centred methods), *or* ii) use performance data of examinees who sat for the assessment and make inferences based on these data (examinee-centred or compromise methods). The best option would be to use *both* ways when setting pass marks, but this is not always possible due to cost and/or other feasibility challenges (Van der Vleuten 2010:175).

Therefore, similar to evaluating and selecting an assessment instrument based on the construct one intends to measure (knowledge, skills or attitude) and the perceived utility of the relevant instrument (Van der Vleuten 1996:54), educators also need to consider the *utility of a standard setting method* before selecting one for implementation. The literature contains a plethora of terms used to describe and/or discuss the critical aspects of standard setting methods. Table 2.3 attempts to simplify the process of evaluating and selecting the appropriate standard setting methods, by organising the key variables (*utility parameters*) into a user-friendly framework, structured into three overarching themes namely:

1. **Robust** (defensibility parameters);
2. **Responsible** (fairness parameters); and
3. **Realistic** (practicability parameters).

TABLE 2.3: UTILITY PARAMETERS TO CONSIDER FOR EVALUATING AND SELECTING A STANDARD SETTING METHOD

Theme	Category	Parameter	Reference source	Comment / Consideration
Robust (Defensibility)	Acceptability	Failure rates	Norcini & Shea 1997:46 Barman 2008:959 Downing *et al.* 2006:57 Kane 1994:432-449	• Method must yield acceptable/realistic failure rates • which are acceptable to educators and regulators
		Objectivity/ Explicability	Bandaranayake 2008:842 Norcini & Shea 1997:41 Van der Vleuten 2010:175 Barman 2008:958 Kane 1994:432-449 Berk 1986:144	• Appropriate panel selection process is vital. Subjective biases of judges must be reduced to a minimum. • How *understandable/explicable* is the method to stakeholders, incl. candidates and lay people.
	Credibility	Validity	Van der Vleuten 2010:175 Norcini & Shea 1997:43 Barman 2008:958 Kane 1994:432-449 Berk 1986:140-144	• External validity correlation (correlation between predicted vs. actual difficulty of items). • Methodology and rationale of method must be sound, evidence-based. • Validation evidence is vital to justify methodology, must be *sensitive to difficulty of test* and stable between similar cohorts and tests. • External markers of competency of passing candidates important.
		Reliability	Kane 1994:445 Peterson *et al.* 2011:7	• The *capacity to reproduce* the pass mark - essential aspect of a method if used in high-stakes assessment. • Internal consistency important for quality assurance.
Responsible (Fairness)	Defensibility	see above	see above	see above
	Transparency	Accessibility	Searle 2000:366 Berk 1986:144	• The availability of information about the standard setting method used to *all stakeholders*
Realistic (Practicability)	Feasibility & Sustainability	Resources required	Van der Vleuten 2010:174 Barman 2008:958 Berk 1986:143 Hambleton & Pitoniak 2006:439	• Affordability of chosen method is critical. • Must be implementable in local context with the available resources (human, time, financial)

In the following section, the terms used in Table 2.3 are described using the published literature pertaining to standard setting.

Producing a valid pass mark

In 1994, Michael Kane published a seminal paper providing a framework for validating the pass mark derived from a standard setting method (Kane 1994:425-457). He explained that since all standard setting methods involve *human judgement* to convert the envisioned, but abstract, "performance standard" into the actual operational pass mark or "passing score", all pass marks are essentially *policy decisions* by educators (Kane 1994:426). Therefore, since there will never be a *gold standard* method or pass mark, it renders all pass marks *arbitrary* (Kane 1994:426; Norcini & Shea 1997:40). However, he explains that although all pass marks are arbitrary, some methods produce pass marks that *are more arbitrary than others*, rendering them indefensible (Kane 1994:426).

The methods that are deemed more arbitrary are the ones i) without a clear policy and rationale which *links* the performance standard to the passing score and ii) where the passing score is set inappropriately for the specified performance standard (Kane 1994:426). Kane explains the above by referring to the traditional fixed 70% pass mark, which was used in the USA as an example of a *more* arbitrary method.

Similar to providing evidence about the validity of the test scores of a particular assessment, pass marks also need to be validated for their intended use (Kane 1994:432). Kane (1994:432), similar to Messick (Messick 1990:15), refers to this principle and the *evidential basis of validity*. This evidence will strengthen confidence in the pass mark, or in its absence or weakness, render the pass mark invalid. Given the arbitrary nature of pass marks, the validity evidence needs to show that administrators and educations have done their best to ensure the pass mark they use to make pass/fail decisions, is appropriate (Kane 1994:435,437) and reasonable (fair) (Kane 1994:434,437).

Deriving an appropriate pass mark

Kane explains that there are two important assumptions that must be satisfied regarding the *appropriateness* of the pass mark (Kane 1994:435). They are the:

- *Descriptive* assumption – the pass mark corresponds to the performance standard. Therefore, candidates on or above the pass mark, have most likely reached the envisioned performance standard and those below the pass mark have not.
- *Policy* assumption – the performance standard is appropriate for the purpose of the pass/fail decision.

Deriving a reasonable pass mark

Reasonableness of the pass mark reflects the impact of the pass mark on pass and failure rates (Kane 1994:434). This is closely linked to the *social accountability* of standard setting, which was explained previously (section 2.2.11) in this chapter.

In gathering validation evidence regarding the appropriateness and reasonableness of the outcome of a standard setting method, Kane proposed the evidence should come from three sources (Kane 1994:437-455):

1. *Procedural* evidence – this refers to the documented systematic process used to derive the passing score.
2. *Internal validity* evidence – this source of evidence refers to the validation of the internal processes used in the standard setting method, such as precision and consistency of item difficulty estimates by the judges. It links closely to the likelihood of the same findings being reached if the process is repeated.
3. *External validity* evidence – these are external validating information that provides evidence that the pass standard generated by the method does in fact lead to pass/fail decisions that would also be found by another method. This is the reason behind the vast amount of research on comparing the outcomes of different standard setting methods, similar to this present study.

Evidence of appropriateness and reasonableness are therefore, inherently linked to the *defensibility* of the standard setting method used as well as the subsequent *credibility* and *acceptability* of the performance outcomes (the pass mark and the resultant pass/fail decisions) (Hobma *et al.* 2004:1245).

Defensibility can be defined from the literature as the robustness of a standard setting method when facing potential challenges to its credibility of converting the envisioned performance standard to an acceptable pass mark (Barman 2008:960; Downing *et al.* 2006:57).

Credibility is defined as the sum of the validity evidence of a standard setting method, as derived by the three sources of validity evidence described by Kane (1994) (Norcini & Guille 2002:814). The particular standard can never be completely validated, but the validity evidence gathered supports its credibility (Barman 2008:960; Norcini & Shea 1997:40). Methods that are supported by research evidence enhance their credibility (Norcini & Guille 2002:817; Norcini & Shea 1997:45)

Acceptability is defined as the extent to which the relevant stakeholders find the outcome of a standard setting method acceptable (Downing *et al.* 2006:57; MacCann & Stanley 2010:143; Norcini & Shea 1997:46). The *objectivity* of the standard setting method's process and the resultant *failure rates* are important parameters determining acceptability (Hambleton & Pitoniak 2006:462).

Objectivity relates to the *explicitness* of the standard setting method's procedures (Hambleton & Pitoniak 2006:457). For example, how this judges were selected (Bandaranayake 2008:842) as well as the sensitivity to influence of bias from individual judges on the performance of the standard setting method.

Explicability is defined as the *ease* with which the standard setting method's rationale and methodology can be explained and understood by stakeholders (Barman 2008:958; Berk 1986:144; Norcini & Guille 2002:818; Schuwirth & Van der Vleuten 2010:205; Van der Vleuten 2010:175). A high level of complexity reduces the explicability of a standard setting method.

Transparency refers to the availability or accessibility of information about the standard setting method and process to stakeholders such as examiners, candidates, patients and any other interested parties (Berk 1986:144).

Feasibility and Sustainability refers to the practicability of implementing, computing and interpreting a particular standard setting method in a local context within the constraints of the available resources (Berk 1986:144; Kane 1998:143; Norcini & Guille 2002:818). This is a critical factor to be considered when selecting suitable standard setting methods and will be discussed in the context of the South Africa in the second part of this chapter (section 2.3.3).

In Chapter 8, the Angoff and Cohen methods of standard setting, the two methods evaluated as part of this study and reported in this thesis, are compared using the utility parameters (UPs) included in the framework shown in Table 2.3, based on their application and performance in this study.

2.3 CONTEXTUALISATION OF STANDARD SETTING IN MEDICAL EDUCATION ASSESSMENT

This part of the chapter explains the context within which this study took place and where this thesis is positioned.

2.3.1 International perspective and impact of standard setting in medical education

Standard setting has been described in the literature for more than 50 years. Nedelsky published his method, specifically intended for use with MCQ tests, in 1954 (*cf.* Nedelsky 1954). Since then, the role and purpose of standard setting in the international *medical* education assessment context has gained considerable momentum in the literature, with many review papers written about the principles and application of different methods in various contexts of assessment in medicine (Bandaranayake 2008:836-844; Barman 2008:957-961; Cusimano 1996:S112-118; Norcini & Guille 2002:811-833).

In 2010, criterion-referenced standard setting was highlighted as an important component of competency-based medical education (CBME), especially in the postgraduate context (Iobst, Sherbino, Ten Cate, Richardson, Dath, Swing, Harris, Mungroo, Holmboe & Frank 2010). In this paper, Iobst and colleagues reflected on the international rise of CBME around the world in countries such as the USA, UK, Netherlands, Canada and Australia.

In 2011, a consensus paper was published following a large international meeting on assessment in medical education (Norcini *et al.* 2011:206-214). The authors categorised the criteria for good assessment and highlighted the importance of standard setting as part of the process of good assessment in medical education (Norcini *et al.* 2011:208).

In 2012, the World Federation for Medical Education (WFME) published a revised version of Global Standards for Quality Improvement of Basic Medical Education (Lindgren *et al.* 2012:1-37). This document lays out the WFME's global standards on two levels, i) the basic or minimum standards expected, and ii) the desired standards for quality improvement in medical education (Lindgren *et al.* 2012:16). In this publication, one of the basic standards expected of a medical school is that it *must* "define, state and publish the principles, methods and practices used for assessment of its students, including the criteria for setting pass marks, grade boundaries and number of allowed retakes" (Lindgren *et al.* 2012:25).

Internationally, many countries have adopted standard setting into their regulatory frameworks for both undergraduate (Clauser *et al.* 2009b; GMC 2009:390) and postgraduate (GMC 2010:66; Plake 1998) medical education and training, recognising the key role of standard setting in the education and assessment process (cf. McManus *et al.* 2014:2). At an international meeting of the Institute of International Medical Education (IIME) in 2005, a committee of IIME went through a pilot process to set the passing standards for graduating doctors internationally (Stern *et al.* 2005:207). With regard to postgraduate medical education and training, the role of standard setting in the certification processes of specialist physicians around the world has also been described in the literature. These are briefly described next.

In the USA, the American Board of Internal Medicine (ABIM) published a paper almost 40 years ago, which describes the use of a norm-referenced standard setting system, based on the field test performance of practising ABIM-certified physicians (Meskauskas & Webster 1975:577). Since then the ABIM changed its standard setting method and currently uses the Yes/No Angoff method in setting passing standards for its certification examinations (Tormey 2014:8).

In the UK, the General Medical Council (GMC), the regulatory body for undergraduate and postgraduate medical education and training, explicitly stipulates the use of standard setting to determine the pass mark for assessments in postgraduate (GMC 2010:15) and undergraduate medical education (GMC 2009:59-60). The Royal College of Physicians (RCP) in the UK, the equivalent body to the CoP in South Africa, used pure norm-referencing (only passing the top 35% of candidates) from 1985 until 2001 (McManus *et al.* 2005:2). In 2002 they implemented a hybrid model, combining elements of the Angoff and Hofstee methods, which was used until an Item Response

Theory (IRT)-based, statistical equating method was introduced for the Part 1 (2008) and Part II (2010), written MRCP (UK) membership examinations (McManus *et al.* 2014:1).

In 1996, the Medical Council of Canada published a paper describing their approach to setting pass marks in their undergraduate or primary medical certification clinical examinations (Reznick, Blackmore, Dauphinee, Rothman & Smee 1996). They introduced the modified Angoff method to set the pass standard for their Objectively Structured Clinical Examinations (OSCEs) in 1993, followed by the introduction of the Borderline group (BG) method for the OSCEs in 1994.

In 2005 a paper from Australia argued for greater coordination and governance of the "fragmented landscape" of Australian postgraduate medical education (Dowton, Stokes, Rawstron, Pogson & Brown 2005). They reviewed the governance and exit-level standard setting systems from other countries and described what Australia could learn from these practices.

In Malaysia, Hassan (2011:1-5) raised his concerns regarding unreliable and invalid assessment practices in postgraduate certification exams in the surgical disciplines in Malaysia. In addition, he expressed concern about the inherent difficulties that would be encountered in attempts to reform these assessment practices.

Twelve years ago, Hutchinson, Aitken and Hayes (2002:73-91) published an extensive review of the literature (1985-2000) on the validation of assessments used in medical certification (licensing) examinations around the world. At the time, only 55 papers, out of more than 7000 titles and abstracts met the criteria for inclusion in this review (Hutchinson *et al.* 2002:76). They reported that general or family practice-based papers were well represented in the literature, but papers reporting on the validity of hospital-based specialties were a rarity. Also of some concern was that there was not a single paper from Africa, including South Africa, included in this review. The high-stakes nature of postgraduate medical certification examinations and the need for increased attention and scrutiny of these certification examinations were strongly emphasised. Norcini and Shea (1997:39-40) have also commented on the significant influence of specialist certification examinations on the careers and career progression of the candidates undertaking them.

Since this 2002 review paper by Hutchinson *et al.*, a positive development occurred in South Africa, with the publication of an analysis of the reliability of the Fellowship of the College of Physicians of South Africa examinations or FCP (SA) (Burch *et al.* 2008). This paper is discussed later in this chapter (*cf.* section 2.3.5). The standard setting of the written examinations of the FCP (SA) examinations was also investigated as part of this present study.

Therefore, the *quality* of medical specialist certification examinations and their results are increasingly under investigation by regulators and researchers from around the world, including South Africa.

2.3.2 Governance and regulation of higher education in South Africa

In this section, the regulatory framework of higher education in South Africa is discussed. Specific reference is also made to the quality assurance of standards and assessment standards.

In South Africa, higher education is governed and regulated by the Department of Higher Education and Training (DoHET); a department of the government of South Africa. The DoHET was formally constituted after the 2009 national general election, when the previous Department of Education was split into the Department of Basic Education (DBE), which oversees primary and secondary education, and the DoHET (RSA DBE 2014).

The previous Department of Education legislated the National Qualifications Framework (NQF) Act of 2008 (RSA 2008), which propagated the enactment of the *National Qualifications Framework (NQF)* for South Africa. The NQF is a "*comprehensive* system approved by the Minister of Higher Education and Training for the classification, registration, publication and articulation of quality-assured national qualifications" (RSA 2008:4). The same NQF Act of 2008 also reaffirmed the *South African Qualifications Authority (SAQA)*, which was established in 1995, as the overarching statutory body charged with advancing the objectives of the NQF as well as managing its implementation and assuring its functioning (RSA 2008:8; SAQA 2000a:8). Figure 2.8 illustrates the basic regulatory structure of the NQF.

The most recent version of the NQF consists of 10 *levels* of registration for qualifications obtainable in the formal education section in South Africa. The first level includes a basic post-secondary school qualification and ends at level 10, the highest level of a higher education (tertiary) qualification, i.e. doctoral level, for example a Ph.D. qualification (RSA DoHET 2013:7).

To manage the vast spectrum of tertiary qualifications more effectively, the NQF Act of 2008 enacted three sub-divisions of the NQF, referred to as *Sub-Frameworks*, each representing the different sectors of formal education and their respective qualifications (RSA 2008:6). Each of the three Sub-Frameworks are constituted and regulated by separate laws (Acts). The three Sub-Frameworks and their respective Acts are shown in Figure 2.8 (RSA 2008:6).

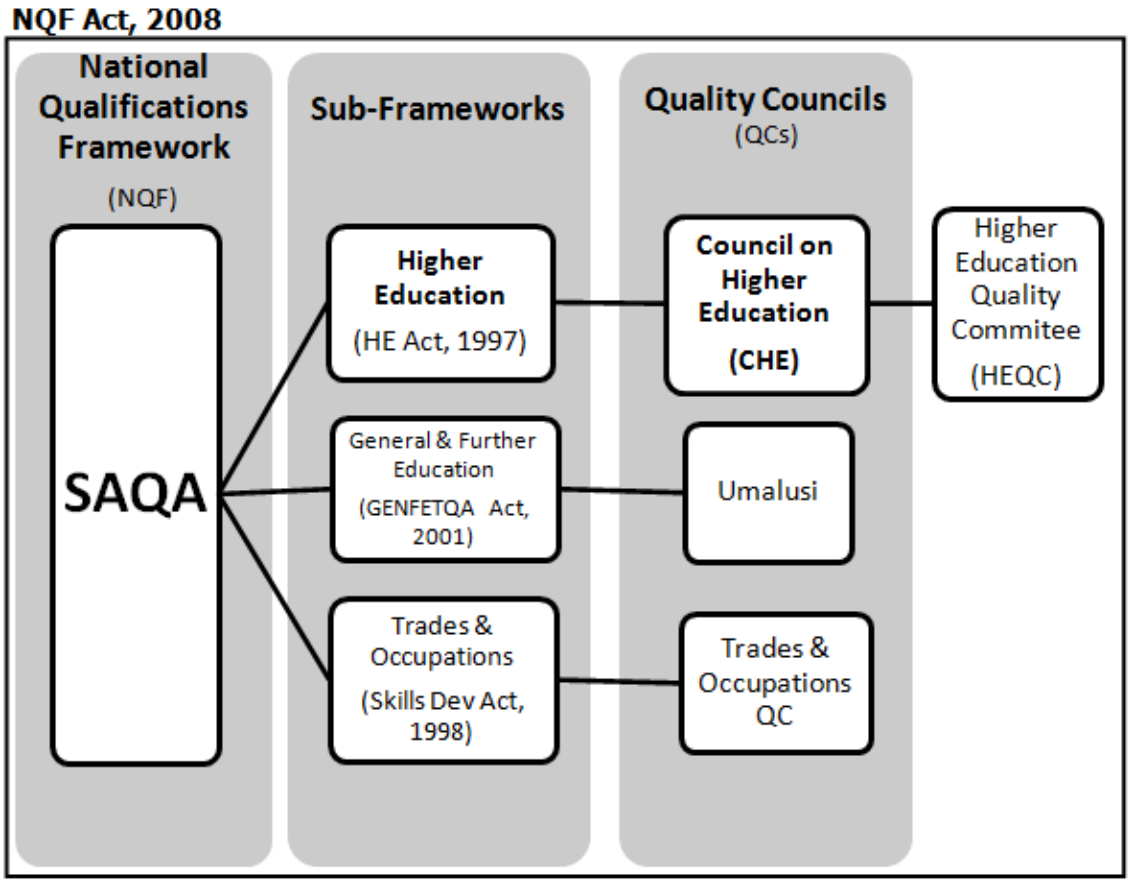


FIGURE 2.8: REGULATORY FRAMEWORK OF HIGHER EDUCATION IN RSA [Compiled by the researcher, SCHOEMAN 2014]

A detailed description of the regulatory framework governing higher education in South Africa is not needed for the purposes of this thesis. However, specific aspects, relevant to the topic of this thesis, namely standard setting, are highlighted.

Each Sub-Framework is overseen by a *Quality Council* (QC), a devolved authority of SAQA, which reports back to SAQA and the Minister of the DoHET about activities under its relevant Sub-Framework.

All higher education qualifications in South Africa, including medical qualifications, must be registered and accredited by SAQA, on the appropriate NQF level. Undergraduate medical degrees, offered by the eight medical schools in South Africa, are registered on level 8, postgraduate medical qualifications are on level 9 and doctoral degrees are on level 10, the highest level of qualification in South Africa (RSA DoHET 2013:7).

SAQA states in their policy documents that the assessment of learners needs to be credible and defines a credible assessment as being a fair, valid, reliable and practicable test (SAQA 2001:16-19). SAQA also specifically defines and mentions the importance of setting defensible and credible standards for the assessments in the learning programmes (SAQA 2000b:15). As a result, the HEQC, on behalf of the CHE, conducts regular accreditation reviews of all undergraduate and postgraduate medical programmes in South Africa, through a paper-based submission system, from a *higher education* perspective.

It is at this point of accreditation and quality assurance in South Africa, where the regulatory process of higher education intersects with that of medical education. The NQF Act of 2008 states that QCs must co-operate with professional bodies regarding the quality assurance of qualifications in their respective occupations (RSA 2008:18). In the next section, the regulation of medical education in South Africa by medicine's own professional regulatory body, the Health Professions Council of South Africa (HPCSA), is discussed.

2.3.3 Governance and quality assurance of medical education in South Africa

In 1997, the Health Professions Act of 1974 was amended to constitute the formation of a new parastatal statutory body, the Health Professions Council of South Africa (HPCSA) (RSA 2009). The HPCSA replaced the previous South African Medical and Dental Council, which was formed in 1928 (Burch 2007:78). The founding of the HPCSA took place as part of significant healthcare reforms that were initiated after the first democratic elections in 1994 and the formation of a new government for the

Republic of South Africa (RSA). The HPCSA, consisting of 12 professional boards representing the full spectrum of health professions registered in South Africa, oversees and administers the annual registration and renewal of registration of all health professionals in South Africa (RSA 2009). It is the equivalent of the General Medical Council (GMC) in the UK or the Medical Council of Canada (MCC).

The Medical and Dental Professional Board (MDPB), one of the boards of the HPCSA, has the mandate to oversee all the medical and dental training in South Africa. The MDPB has various committees and sub-committees dealing with medical education and assessment. The two most important committees, from an accreditation and quality assurance in medical education perspective, are the Sub-committee for Undergraduate Education and Training, as well as the Sub-committee for Postgraduate Education and Training (Medical). These important structures are graphically illustrated in Figure 2.9.

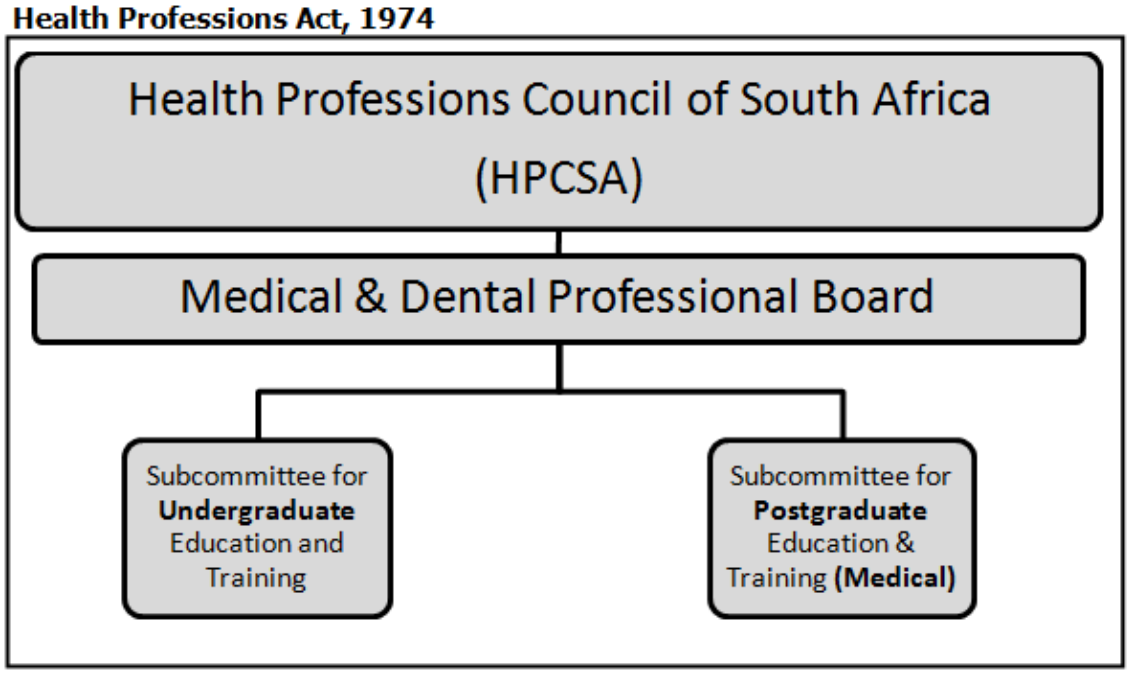


FIGURE 2.9: REGULATORY FRAMEWORK OF MEDICAL EDUCATION IN RSA [Compiled by the researcher, SCHOEMAN 2014]

The Sub-committees for undergraduate and postgraduate education and training of the HPCSA are accredited by SAQA (SAQA 2012), via the CHE, and they conduct 5-yearly accreditation visits to all undergraduate and postgraduate medical programmes in South Africa. The HPCSA's accreditation process focuses on quality assurance, of the respective medical training programmes, from a *medical* education (occupational) perspective. This means that all medical programmes in South Africa undergo two

independent accreditation processes conducted by i) the CHE (previously described) and ii) the HPCSA.

Previous research on medical education and assessment in South Africa

A comprehensive overview of assessment practices in medical education in South Africa, was conducted by Vanessa Burch, a specialist physician and Professor of Clinical Medicine at the University of Cape Town in South Africa, as part of her Ph.D. thesis published in 2007 (Burch 2007). She investigated the topic of assessment as it relates to the training of medical doctors in South Africa. The foundations of her Ph.D. work and research centred around the *purposes of assessment* in the training of doctors, as well as the 'utility' of assessment instruments in resource-limited settings typical of developing countries, like South Africa. Burch used and expanded on Van der Vleuten's (1996) description and definition of the 'utility' of a given assessment instrument, to describe her work in the field of medical education assessment in South Africa. Van der Vleuten described the utility of an assessment instrument is determined by the rigour (validity and reliability), educational impact and practicability (acceptability of stakeholders and the cost of administering the assessment) of the specific assessment tool (Van der Vleuten 1996:54).

Part of the work done by Burch (2007:129) included a comprehensive evaluation of the reliability of the postgraduate certification examination, the FCP (SA), of the CoP in South Africa. She used generalisability theory to assess the reliability of the different components of the FCP (SA) examinations. The findings of this research were published in 2008 (Burch *et al.* 2008) and served as a foundation for the work reported in this thesis, which is discussed in more detail later in this chapter (section 2.3.5).

In her thesis, Burch provides a detailed account of the relationship between healthcare resources and the burden of disease in South Africa (Burch 2007:81-97). She explains that although South Africa is the wealthiest country in Africa and has the most doctors per 10 000 of the population in Africa, only 37% of the doctors in South Africa work in the public sector, which serves 85% of the population. This reduces South Africa to a doctor:population ratio reported in some of the poorest countries in Africa. In addition, she makes the point that medical education, training and assessment happen almost exclusively in the public health sector, and so the task of supervising and assessing undergraduate medical students and postgraduate specialist trainees falls virtually entirely on the clinician-educators employed in understaffed and over-

burdened public healthcare sector. This critical shortage of clinical-educators in South Africa is a significant limiting factor when new initiatives in medical education assessment, including standard setting, are considered for implementation, as is the case in the work presented in this thesis.

The current study builds on the work done by Burch (2007) by investigating one of the purposes of assessment, which is to make decisions about students' progression in training and certification (i.e. pass/fail decisions) as well as the processes involved in coming to these important decisions (standard setting). This study and thesis focuses on progression and certification decisions made about postgraduate medical (specialist physician) trainees in South Africa.

2.3.4 Governance and regulation of postgraduate (specialist) certification examinations in South Africa

In 2005, postgraduate medical education and training in Australia was described as a "fragmented landscape" (Dowton *et al.* 2005:177). The authors reflected on the lack of structure and overall governance of the interactions between the different stakeholders involved in the postgraduate medical education and training processes. Dowton and colleagues encouraged the relevant regulatory bodies and employers to review postgraduate training and employment models in other countries such as the USA, Canada and New Zealand and learn from the lessons of these countries where the postgraduate specialist training 'landscapes' had been reformed (Dowton *et al.* 2005:177-180). Hays (2007:400-403) provided some additional insight by warning Australasian countries, contemplating reform, of the painful and negative effects of rushing through radical changes in a large and established medical education system. He described the challenges experienced in the UK during the Modernising Medical Careers (MMC) reform process of postgraduate medical education and career selection system (Hays 2007:400-403).

In South Africa a similar process of postgraduate medical education reform has recently begun. As explained in Chapter 1, up to 2010, a complex *parallel* examination and certification route existed for trainees registering as medical specialists with the HPCSA. In 2010, this dual system entered a change process as the Sub-committee for Postgraduate Education & Training (Medical) of the HPCSA made a decision (HPCSA 2010) to centralise the summative *exit* examinations for all trainees who commenced their specialist training on or after 1 January 2011 (HPCSA 2011).

The HPCSA and CMSA recently signed a memorandum of understanding (MoU) to formalise the mandate the CMSA had received from the HPCSA to conduct the exit licensing examinations for all the medical and dental specialities in South Africa (HPCSA & CMSA 2014). Since the CMSA exit examinations are now the only route to registering as a specialist in South Africa, it has placed a significant responsibility on the CMSA to ensure that their examinations are rigorous and fair to candidates. As discussed in this chapter and throughout this thesis, standard setting is a critical component of the quality assurance process of summative assessments because the results and subsequent judgement decisions made about candidates have far reaching implications. Now, more than ever, the mandate for rigorous, quality assured assessment processes, including standard setting, in the CMSA examinations is clear.

The topics of assessment reform, improvements and standard setting are not new discussions in the CMSA. The CMSA held a symposium on postgraduate assessment in May 2003 to review and discuss a strategy and vision for revising and improving on the quality of its assessments and assessment processes. In a subsequent report, Hift & Burch (2003:76-77) list numerous quality improvement strategies to be addressed, including i) explicitly defining the curricula and learning outcomes for each College, ii) using psychometrically sound and credible assessment instruments and iii) defining the passing thresholds for assessments in the respective Colleges more rigorously by using either criterion- or norm-referenced standard setting methods (Hift & Burch 2003:76).

Kent (2003:78-79), in his commentary on the CMSA assessment symposium held in 2003, highlighted that the CMSA needed to regularly review and improve its assessment systems, in keeping with the importance and high-stakes of its examinations and the impact they have on the candidates undertaking them. He also made the point that high failure rates of CMSA examinations are "educationally unsound" and not supported by the medical education literature as a sign of "maintaining standards" (Kent 2003:78). He suggested that it was more a sign that the "educational system is at fault" (Kent 2003:78). He concluded by making a call for the respective member Colleges to work together as far as possible to share best practices and good innovations in educational practice.

At the time of writing this review the CMSA did not have publicly available (website or hard copy publication) policies or regulations regarding standard setting in the assessment processes of its member Colleges (CMSA 2015:online). To date, only the

CoP has implemented the use of a methodological process to determine the pass marks of its written examinations; all the other CMSA member Colleges still use the fixed 50% pass mark (CMSA 2015:online; CoP 2013:4).

The process of introducing standard setting into the CoP was briefly explained in Chapter 1. In the next section of this chapter, previously published research on assessment practices in the CoP from 2001 – 2005 is reviewed and discussed. This work is described in detail because it contributed to the introduction of standard setting in the CoP, as well as providing additional motivation for the work presented in this thesis.

2.3.5 Assessment and pass standards for postgraduate specialist physician training in South Africa

The FCP (SA) examination processes and results have previously been studied and reported in two papers by Burch and colleagues (Burch & Norman 2009; Burch *et al.* 2008). Some of this work originated from her PhD research (Burch *et al.* 2008) and the other paper from work done thereafter (Burch & Norman 2009). These two papers are particularly relevant for this thesis, since they are the only directly related papers, in terms of assessment context, and provide an accurate historical perspective and background to the present study. Many of the findings reported by Burch *et al.* (2008) and Burch & Norman (2009) are important to consider when interpreting the results of the present study.

Burch concluded both papers with strong and specific calls for further research on standard setting, particularly in the context of high-stakes postgraduate licensing examinations (Burch & Norman 2009:446; Burch *et al.* 2008:532). These calls contributed significantly to the momentum required to launch the work reported in this thesis.

Both papers focused exclusively on the *Part II* (exit-level) components (written and clinical) of the FCP (SA) examination (Burch & Norman 2009:443; Burch *et al.* 2008:524). Part I data were not reported in either paper. The results from the two papers have been synthesised to provide an overview of key findings (*cf.* Table 2.4). This was possible as they relate to two consecutive historical timeframes of the FCP (SA) Part II examination: 2001 to 2003 (Burch & Norman 2009:443) and 2004 to 2005 (Burch *et al.* 2008:524). This overall time frame (2001-2005) contained ten cycles of

the FCP (SA) examination data, two per calendar year. The data from all of these cycles, except May 2004 were included in the two papers. Reasons for excluding the May 2004 were not provided by the authors, but it was the first cycle where the new numerical scoring system was used for the written and clinical components, as opposed to the old letter-based grading system (Burch & Norman 2009:443).

It is clear from reading the two papers that the 2001-2003 data were used to calibrate and justify the conversion from a letter-based grading system to a numerical system, using a fixed 50% pass mark. The data from the 2001-2003 paper served to reassure examiners that the grading/scoring systems were comparable (Burch & Norman 2009:445).

This section of the chapter focuses on the data which describes the *written* components of the assessment process, since it is directly relevant to the present study and this thesis.

Table 2.4 shows that the total number of candidates entering the FCP (SA) Part II examination across the nine cycles was 220, with an average of 24 candidates per cycle. The overall mean failure rate of the composite written component (SAQ and DIT) was 21.4% and the clinical component was 27.2%. This resulted in an overall mean failure rate of 42.7% for the FCP Part II exam over the nine cycles from 2001 - 2005.

TABLE 2.4: COMBINED DATA FROM PREVIOUS FCP (SA) PART II EXAMINATIONS

Sat written papers	**2001 – 2003** 6 cycles **n=141**	**2004 – 2005** 3 cycles **n=79**	**2001 – 2005** 9 cycles **n=220**
Failed written (%)	37 (26.2)	10 (12.7)	**47 (21.4)**
Passed written (%)	104 (73.8)	69 (87.3)	**173 (78.6)**
Invited to clinical	**n=104**	**n=69**	**n=173**
Failed clinical (%)	32 (30.8)	15 (21.7)	**47 (27.2)**
Passed clinical (%)	72 (69.2)	54 (78.3)	**126 (72.8)**
Final outcome			
Failed Overall (%)	69 (48.9)	25 (31.6)	**94 (42.7)**
Passed Overall (%)	72 (51.1)	54 (68.4)	**126 (57.3)**

It can be seen from Table 2.4 that there was a reduction in failure rates in the written *and* clinical components from the previous letter-grading system (pre-2004) to the newer numerical-scoring system with a clear 50% pass mark (2004 and beyond). Reasons for the reduced failure rates are not known, but better candidates, a more lenient assessment scoring system, a combination of both or a clearly defined pass mark may have contributed.

It is important to note that during the time of these two studies (2001-2005), the written component was a *composite* examination comprising of the marks from the theory papers (SAQ tests) and a Data Interpreting (DI) test. The pass/fail decision was based on the mean combined mark from these two written components. In the present study, candidates must pass each written component *independently* (conjunctive standard) (CoP 2011b:5; McKinley & Norcini 2014:108) and this could result in higher failure rates (McKinley & Norcini 2014:109). Similar to the present study, *only* the candidates who passed the written component were invited to the clinical component of the Part II examination. The SAQ tests in these Burch-papers are equivalent to the SEQ tests of the present study and the DI-test (Burch-papers) to the Objective Test of the present study.

The reliability of the Part II examinations were calculated using generalisability theory for the three examination cycles between 2004 – 2005 (Burch *et al.* 2008:527). The mean data across the three cycles were used and the reliability of the Part II examination was calculated as 0.72, with a SEM of 4.17%. The authors only used the data of the candidates who passed the written papers (scored 50% or more for the written component, n=69, *cf*. Table 2.4) and hence were invited to the clinical component of the Part II examination. This was understandable since they wanted to get an idea of the reliability of the *whole* Part II examination, but it does limit the interpretation of the data, since only the strongest candidates' data were in effect used. As a result, the SDs of the written data were quite narrow (*cf*. Table 2.5 below) - 5.4% and 8.8% for the SAQ and DI test respectively (Burch *et al.* 2008:524). That probably contributed to the low reliability measurements of 0.59 (SAQ) and 0.64 (DI test) of the individual written components. The SEMs, which control for SDs, were 4.45% (SAQ) and 7.02% (DI test) respectively and provide a sharper estimate of the accuracy of these criterion-referenced written tests. The *composite* written component reliability was not reported, but would probably have been above 0.64 (the highest

reliability of one of the components). The reliability of the written papers in the 2001-2003 dataset was not reported.

TABLE 2.5: DESCRIPTIVE STATISTICS OF PREVIOUS PART II WRITTEN COMPONENTS (2001 – 2005)

Descriptor	**2001 – 2003** 6 cycles **n=141**	**2004 – 2005** 3 cycles **n=69***
Mean of **SAQ** (SD) %	55.4 (12.4)	58.7 (5.4)
Mean of **DI test** (SD) %	46.8 (24.4)	56.8 (8.8)

*Only data of candidates who *passed* the written component reported in the paper

Table 2.5 provides the mean scores (and SDs) for the 2001-2003 data, of *all* the candidates (passed or failed, n=141) who sat the written tests after the conversion from a letter-based grading system to numerical values (Burch & Norman 2009:445). The SAQ mean score was 55.4% (SD= 12.4%) and the DI test mean score was 46.8% (SD= 24.4%). This was in contrast to the 2004-2005 data (only the candidates who passed the written component) where the SAQ mean was 58.7% with a 5.4% SD and the DI test was 56.8% with a SD of 8.8% (*cf.* Table 2.5).

The mean SAQ mark was similar for the two datasets, but the SD is more than halved in the 2004-2005 dataset (reflecting the narrower ability range of the more selected candidates who passed the written component, *cf.* Table 2.5). The DI test data in the 2001-2003 (all data) cohort had an 8.6% lower mean than the SAQ data, with double the SD (24.4%). The DI test was considerably more difficult and more discriminating than the SAQ in this dataset. This was not the case in the 2004-2005 dataset, where the means and SDs of the two test components were much more aligned, reflecting again the characteristics of a more homogenous, higher ability cohort, who all passed the written papers.

The low number of test items, 28 in total, used in the written papers (SAQ and DI test) was reported and evident in both Burch-papers. This cast doubt on the representativeness and subsequent generalisability (validity) of the results. It almost certainly contributed to the low reliability measurements of the written tests. No item analysis data were reported in either of the Burch-papers, which means that

commenting on the quality of the examination results, from a psychometric perspective, is not possible.

Current regulations pertaining to the use of standard setting in the CoP

As previously mentioned in Chapter 1, the CoP introduced the Cohen method of standard setting (Cohen-Schotanus & Van der Vleuten 2010) for its written entry-level FCP (SA) Part I MCQ test, in 2011 (CoP 2011b:4). In 2012, the Cohen method was also introduced for the written exit-level FCP (SA) Part II Objective test (CoP 2011b:5) and in 2013 for the FCP (SA) Part II SEQ (CoP 2013:4).

The next section discusses the *context* in which standard setting was introduced in the CoP - specifically the 'hearts and minds' of the examiners.

2.3.6 CoP examiners' knowledge, attitudes and perspectives on standard setting

In South Africa, the concept of standard setting is relatively new and awareness about the use of formal standard setting processes is lacking (Pitoniak & Yeld 2013:23). A recent standard setting study, conducted in South Africa by Pitoniak and Yeld (2013), investigated the thoughts and views of standard setting panellists about their experiences of taking part in an Angoff standard setting process for the first time. They were all part of a standard setting process for a national, pre-university, entrance examination on Academic Literacy, Quantitative Literacy and Mathematics, called the National Benchmarking Tests (NBT) (Pitoniak & Yeld 2013:24-26). The study reported that many panellists had fears about the effect these 'cut-scores' would have on the higher education opportunities of students sitting these tests, given the persisting inequity of the South African, pre-university, education system. The political and educational impact of criterion-based standard setting emerged as powerful variables to consider, when designing future standard setting studies. In addition, the appropriate representative composition of panels was emphasised as a very important consideration when selecting standard setting panellists (Pitoniak & Yeld 2013:29).

Standard setting is slowly gaining traction in undergraduate and postgraduate medical education. Introduction of the Cohen method in 2011 in the FCP (SA) written examinations, the first College in the CMSA to introduce a formal standard setting

methodology other than the fixed 50% pass mark (CMSA 2015:online), was a landmark event for postgraduate medical education assessment in the country.

Not unlike the experience of Pitoniak and Yeld (2013), the introduction of standard setting in the CoP occurred in a context where very few examiners of the CoP had prior knowledge, exposure or experience in standard setting. Indeed, their actual knowledge, views, attitudes and perspectives about standard setting were not known at the start of this study. In order to effectively guide and introduce such a significant change in assessment practice and pass/fail decision-making policies in the CoP, the researcher and supervisors expanded the study to include a situational analysis of the CoP examiners' knowledge, views, attitudes and perspectives at the start of the study and towards the end. The information gathered from the first situational analysis was used to develop an educational seminar for the examiners about the concept of standard setting and in particular how the Angoff and Cohen methods worked.

The results of the change management process (situational analyses and educational seminar) and how the CoP examiners' knowledge, views, attitudes and perspectives on standard setting changed over the course of the study are presented in Chapters 4 and 8.

As part of the literature review of this thesis, a search was conducted to identify published work reporting on the knowledge, views, attitudes and perspectives of examiners regarding the *introduction* of standard setting for an examination or examination system. All the studies identified in the literature about the views of standard setting panellists (the judges) related to the actual standard setting process they were exposed to, their views about it and the resulting standards that were set (Cizek *et al.* 2004:45; cf. Hambleton & Pitoniak 2006:456; Hansen *et al.* 2013; Impara & Plake 1997). This approach was also followed in a recent study where the Angoff standard setting process was conducted in a similar context to the present study – in sub-Saharan Africa (specific country not reported). The authors were interested to explore if the panel in Africa (different culture) had similar or divergent thoughts and views on the standard setting process as those in the USA (Ferdous & Buckendahl 2013). Their results reported some similarities and some differences to American standard setting panels.

The reasons for this gap in the literature are speculative, but one possible reason might be that in most countries standard setting has become a mandatory part of the regulatory systems of general education and medical education, and so the research question about what examiners (or judges) know or how they feel about standard setting and the introduction of standard setting is no longer relevant. Another reason, perhaps the other extreme, is where standard setting is not used in a community of practice and the status quo of setting pass marks, in whichever way, is maintained and *not questioned*. In this context, there is no perceived need for the topic to be researched. However, between these two extremes there is a clear gap in the literature for educators working in educational systems that are embarking on, or are planning to go through, this big assessment transformation process in which a change management is essential.

To effect change in assessment practices is difficult (Burch & Norman 2009:445). Since a significant portion of the work reported in this study and thesis is about bringing change to assessment practices in the CoP, a better understanding of the *change process* and *change management,* is essential.

Therefore, the next section of this chapter is dedicated to reviewing change theory and change management strategies, with reference, where possible, to medical education. In addition, the way a new (educational) innovation diffuses through an organisation is also reviewed and discussed.

2.3.7 Change management and the diffusion of an innovation

This section of the chapter brings together two components of the literature. One is the different theoretical models of a change management process and the other is the diffusion of an innovation through an assessment organisation. These literature components relate to the scope of the thesis as a result of the change process that took place surrounding the introduction and implementation of a new assessment innovation (standard setting) in a social system (the CoP). These two aspects were closely aligned and influenced one another.

An exhaustive review and discussion of all the possible change management theories and approaches described in the published literature falls outside the remit of this thesis. The focus of this review was to identify and discuss change management

theories that have been linked to change processes as a result of introducing innovations or reforms in medical education. These will be synthesized where there are overlaps and contextualised to this study.

"You do not really understand an organisation or system until you try to change it"
Kurt Lewin (1952)

The above quote by Kurt Lewin from 1947 carries universal truth for anybody who has ever attempted to be a change agent. Change and change management are complex, difficult and dynamic processes (Agius, Willis, McArdle & O'Neill 2008:e87; Gale & Grant 1997:249; Norcini & Banda 2011:83; Prideaux 2004:2). Those individuals attempting to reform and improve human systems and organisations naturally try to effect positive change in them, but many times are subsequently confronted by significant resistance, obstacles or inertia to their change efforts (MacFarlane, Gantley & Murray 2002:320; Mennin & Kaufman 1989:9). Only then do they begin to understand and unravel the hard realities of the forces and undercurrents inside the organisation or system, reflecting the true nature and opinions (organisational culture) of the human 'change targets' on the proposed changes.

In an effort to facilitate more harmonious change processes, many authors have published guidelines and strategies to foster a better understanding in change agents of the underpinning principles and concepts of change and its management, both in general (Schein 1996; 2002) and more specifically in medicine and medical education change management (Gale & Grant 1997; Graves & Burch 2012:1123; Mennin & Kaufman 1989).

One of the commonest models used to describe and understand the process of change in any human system (big or small) is Kurt Lewin's simple 3-stage change model of *unfreezing, changing and refreezing* (Lewin 1952). This model provided the basic framework for the work of Edgar Schein, who developed an expanded, more comprehensive model, explaining the underlying processes informing the change process (Schein 1996; 2002). Figure 2.10 provides a graphic representation of Lewin's change model. This 3-stage model will form the basis for this discussion on change and change management and other change models or strategies described in the literature will be linked at the appropriate stage.

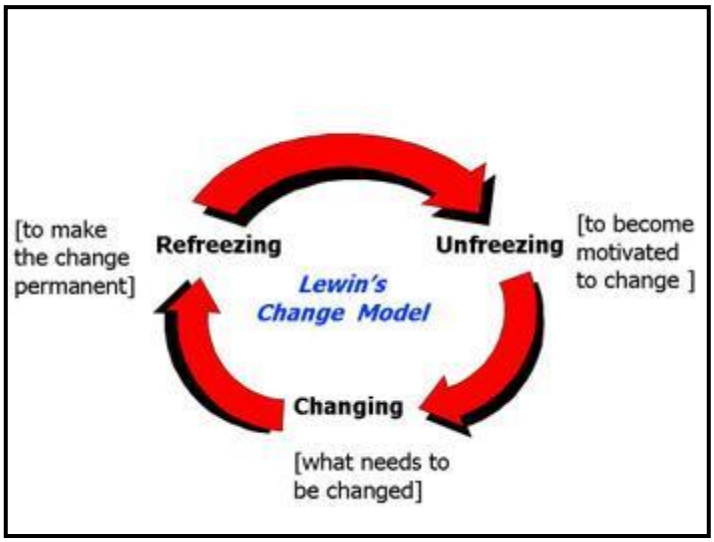


FIGURE 2.10: LEWIN'S CHANGE MODEL
[Figure taken from Google images]

2.3.7.1 *Stage 1 - Unfreezing*

This is the first stage of the change process where the change targets (people) need to become motivated to change their views, attitudes and/or behaviours. Schein describes that all human systems hover in a state of "quasi-stationary equilibrium". This means that although they might appear stable or stationary, they actually are not. There are constant forces of change pressing against opposing forces of resistance to change. When a system does not have strong 'changing' forces at work at a given time, the resisting force is equally low and hence the system remains in its present state. In a sense, this keeps the system (or organisation) 'frozen' and real change cannot be effected. This 'force field' of opposing forces needs to be well understood and analysed if a change agent hopes to be effective (Schein 1996:1-2; 2002:36-38). Figure 2.11 graphically illustrates the force field analysis concept. Schein goes on to explain the sub-stages to unfreeze a system to enable change to occur.

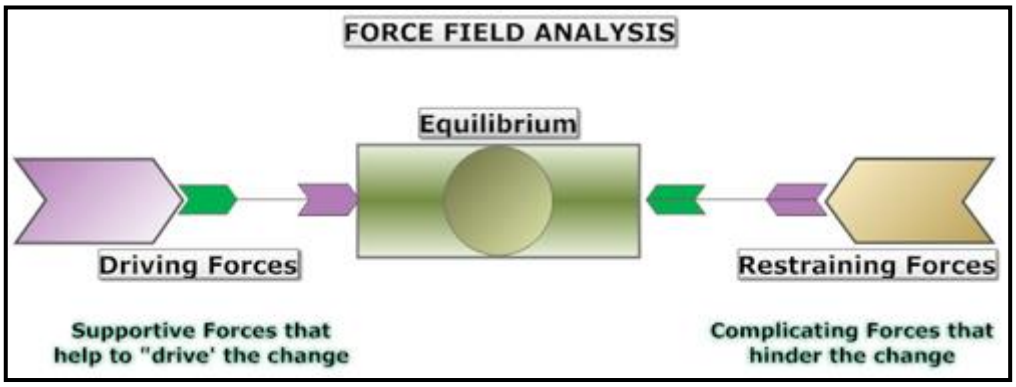


FIGURE 2.11: FORCE FIELD ANALYSIS
[Figure taken from Google images]

Disconfirmation

This is disrupting or disturbing information or events that weaken the resistance to change in a particular system. It provides the motivation to change and initiates the unfreezing process (Schein 1996:1-2; 2002:36-38). In a UK undergraduate medical education case-study on change management after the introduction of a large community-based clinical skills teaching initiative in London, the authors describe the efforts they had to make to change the attitudes and behaviours of five medical schools to enable change (MacFarlane *et al.* 2002:323). They also refer to another dual-option strategy described by Hunt (*cf.* MacFarlane *et al.* 2002), which addresses the changing of attitudes and behaviours in the human system. The one option is 'Involve those affected' – which refers to leaders engaging in a consultation process to develop 'buy-in' in a 'bottom-up' approach to motivate change. This is the preferred approach to use with doctors and in medical education (Gale & Grant 1997:240). The alternative, more authoritarian, option of 'tell those affected' is the 'top-down' approach which is less ideal, but might be applicable when the urgency for change is great and the time available is limited. Prideaux (2004:2) argues that both these approaches have merit, depending on the nature and size of the change and the context and culture where it takes place. Effective strategies to deliver disconfirming messages is by presenting data from the current state, running pilot projects or external reports (Gale & Grant 1997:242; Mennin & Kaufman 1989:14)

Survival anxiety

The disconfirmation data leads to a state of guilt or as Schein puts it "survival anxiety". The stage induces a real discomforting psychological anxiety state where the change target's pervious beliefs and/or behaviours become invalid in their own perspective (Schein 1996:1-2; 2002:36-38).

Gale and Grant (1997:241) refers to this stage as creating the need to change. They stress it is important not to offer solutions too early, i.e. before the problems in the current state have been clearly defined and articulated. A crisis in the system rapidly supports the development of this stage (Mennin & Kaufman 1989:9,14). The perception that the future of the organisation or system is at risk becomes vivid. In the medical education context, this could translate to fears that accreditation might be lost or not retained, pass standards might not be credibly set and litigation could be imminent or the reputation of the institution is slipping.

However, this survival anxiety may not be sufficient to induce change if it triggers enough "learning anxiety" in the members of the system (Schein 1996:1-2; 2002:36-38). This occurs when the system becomes defensive, goes into denial and subsequently avoids addressing the disconfirmation data in fear of losing face or being exposed as incompetent. The pain of 'unlearning' the old ideas or behaviours is greater than the survival anxiety. Therefore, for change to occur, either the survival anxiety must exceed the learning anxiety, or preferably, the learning anxiety should be reduced. The development of learning anxiety is prevented or reduced, when sufficient psychological safety is provided by the change agent.

Psychological safety

This is the stage when the change agent's innovation is regarded as appropriate and safe to engage with. It is the final step that opens the door for change to take place (Schein 1996:1-2; 2002:36-38).

However, Schein explains that it is a regular occurrence for disconfirming information to have existed for a long time in a system but, due to a lack of psychological safety, an organisation or group of individuals will continue to avoid engaging with change by repressing or denying the relevance, validity, or even the existence of the information. The fundamental nature of psychological safety is to generate the conditions that allow the consideration of required change without feeling a loss of integrity or identity (*cf*. Agius *et al.* 2008:e92). Strategies offered by other authors on providing psychological safety include: consult widely on future plans, demonstration projects, gradual implementation strategy, constant review of new innovation (Gale & Grant 1997:242; Mennin & Kaufman 1989:15)

Therefore, in summary, the art of the unfreezing stage of change management lies herein, that the change agent provides enough discomforting information to induce survival anxiety, but also provides sufficient amounts of psychological safety to motivate the change target to rather embrace the change process, as opposed to retract into learning anxiety and denial.

2.3.7.2 *Stage 2 - Changing*

This is the stage where the actual change happens and the new reform or innovation is implemented (Schein 1996:2-3; 2002:38-39). The system has thawed from the frozen

state and is now in a fluid state that moves to a new "quasi-stationary equilibrium" (Gale & Grant 1997:244). The identification of how to implement the reform and the detailed action plan now becomes important (Gale & Grant 1997:246). The process of 'cognitive restricting' occurs during the change phase, where the individuals need to learn and adapt to new ways of thinking and behaviour. In the study conducted by Agius *et al.* (2008:e93) this aspect was still problematic for many of the medical consultants who were involved in the large-scale restructuring of postgraduate medical education career paths in the UK, called Modernising Medical Careers (MMC). Hays, also commenting on the MMC change process, as studied by Aguis *et al.*, concluded that the critical error that was made during the implementation of the MMC changes was managers "moving too fast, too soon" (Hays 2007:403). It seems MMC changes were forced down onto an unfrozen UK medical workforce landscape and hence stakeholders were alienated.

Role-models and good leadership are critical aspects to effect positive change processes (Eccles 1994; Gale & Grant 1997:243; MacFarlane *et al.* 2002:326). Lieff and Albert (2012:317) reports that senior medical education leaders in one Canadian medical school were very aware of their role as providing leadership in change processes.

Grant and Gale (1997:244) also stressed the importance of the appropriate timing of the change process, which relates closely with how the innovation will diffuse through the hearts and minds of the people. A more detailed discussion on the diffusion of an innovation follows later in this section.

2.3.7.3 *Stage 3 - Refreezing*

The final stage of a change process is the refreezing stage (Schein 1996:4-5; 2002:39-41). During this phase the newly changed system stabilises again and the effect of the change becomes embedded in their thoughts. The cognitive restructuring has finished. Now the shift towards owning the changed, new position occurs, from the change agent to the members of the system. Integration and internalisation of the change takes place into the identity of the members (Mennin & Kaufman 1989:15). This is an important step if the new change or innovation is to last (Schein 1996:4-5; 2002:39-41).

The importance of change *leading to improvement* in the system, both from the perspective of the members of the system and their customers (students and patients in the context of medical education) has been stressed by authors (Langley, Moen, Nolan, Nolan, Norman & Provost 2009:3)

Pockets of resistance might remain in the system, but must not be confused with opposition. Qualified acceptance will help to ensure the new changed position achieves its best possible fruition with available resources – it's the enemy of complacency (Gale & Grant 1997:247). The perpetual or continuous nature of change is well recognised and, as mentioned at the start of this section, the change process is complex with change agents operating regularly in all three stages during a change process.

2.3.7.4 *The role of change agents or champions*

As mentioned above, the role and leadership abilities of change agents are paramount in the management of effective change.

Everett Rogers defines a champion (in the context of being the principal change agent for a new innovation in an organisation) as: "A champion is a charismatic individual who throws his or her weight behind an innovation, thus overcoming indifference or resistance that the new innovation may provoke in an organisation" (Rogers 2003:414) The presence of a champion contributes directly to the success of an innovation in an organisation. Their impact and influence tends to be more powerful when the innovation is radical. Naturally, an anti-innovation champion, who opposes the innovation, will have a similar effect, but in the opposite direction (Rogers 2003:414). In terms of organisational change, Schön (1963:84) stated the effect of a champion even more bluntly: "The new idea either finds a champion or dies".

Rogers (2003:415) goes further to describe the characteristics of champions:

1. They occupy key positions in the organisation;
2. Have good analytical and intuitive skills about their colleagues;
3. Have good interpersonal and negotiating skills;
4. They are effective at making an innovation fit into an organisation.

Schein describes them as powerful 'unfreezers' of a human system (Schein 2002:37). They are also influential in how the innovation diffuses through the organisation. In the context of this study, the role of a champion in the CoP was clearly evident and supported the introduction and implementation of standard setting in the CoP (the innovation)

In many contexts, especially in resource constrained ones, the change agent is most likely involved in all three stages of the Lewin's 3-stage model of change (Rogers 2003:369-370; Schein 2002:41). They are the innovators, implementers and need to provide leadership for effective refreezing (Burch 2011:24; Norcini & Banda 2011:85). There is evidence emerging that effective, socially accountable faculty development initiatives are delivering more change agents in resourced constrained regions of the world (Burdick, Friedman & Diserens 2012). However, faculty members are not the only change agents in medical education.

Medical students are sometimes the unexpected, informal change agents, as was the case in one UK study (MacFarlane *et al.* 2002:324). Therefore, equipping them with change management skills seems sound and sensible. The reform of medical curricula to equip medical students as change agents in their local contexts is viewed as an important global social accountability initiative and The Network: Towards Unity for Health is actively promoting this strategy (Christobal, Engel & Talati 2009:4).

One role in change management, which is usually reserved for faculty members or sometimes external consultants, is that of champions of innovation (or reform) within an organisation.

2.3.7.5 *The diffusion of innovations*

The model of how an innovation diffuses through a social system was first described in 1962 by Everett Rogers in his book, Diffusion of innovations (Rogers 2003:39). He defined diffusion in this context as: "The process in which an *innovation* is *communicated* through certain *channels* over time among the members of a *social system*" (Rogers 2003:5).

Rogers describes five types of adopter categories, which reflect the *innovativeness* of the individuals in the social system. The speed at which an individual embraces a new

innovation corresponds to their innovativeness. Therefore, the categories match the adoption rate of individuals in the system (Rogers 2003:280). Innovativeness is a continuous variable, which follows a normal distribution in a given population, similar to height. The five adopter categories, in time order (early to late) after the launch of the innovation, are illustrated in Figure 2.12.

The orange, S-shaped cumulative percentage adoption curve and the blue, bell-shaped distribution curve, with the proportions in each category are shown in Figure 2.12.

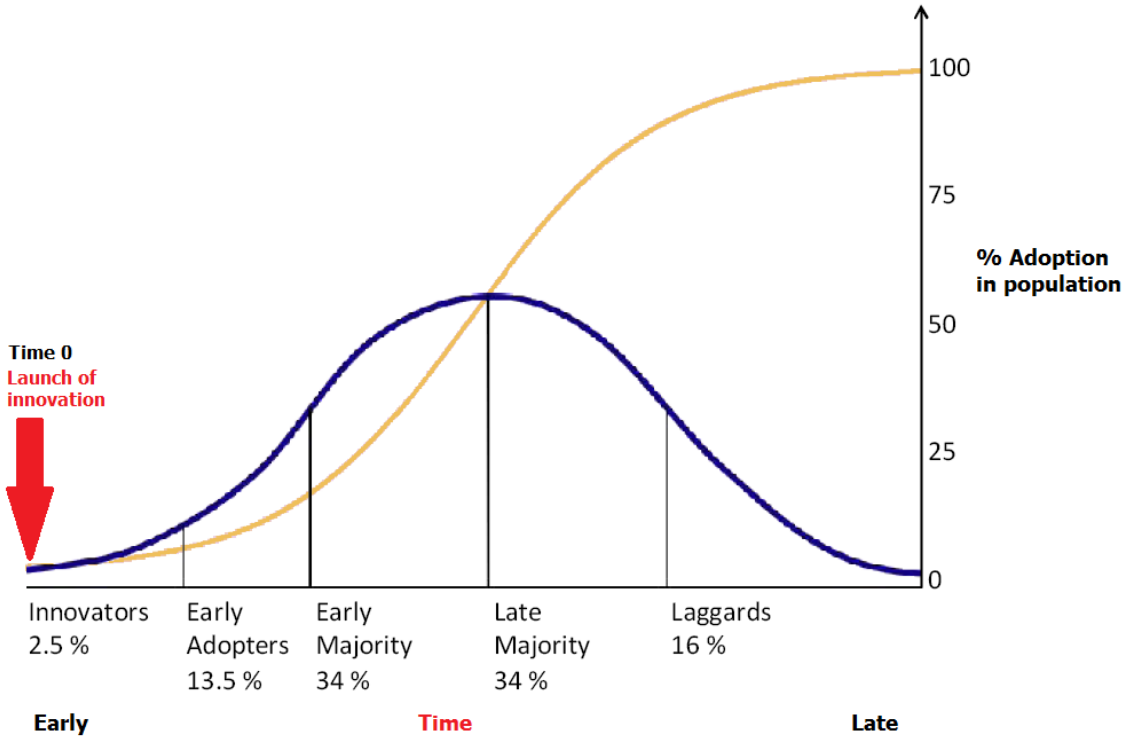


FIGURE 2.12: DIFFUSION OF INNOVATION BY ROGERS (2003:281) [Adapted from Rogers (2003:281) - Time annotation added]

The five adopter categories are not exhaustive of all the individuals in the system. The non-adopters are not illustrated by Figure 2.12. They represent the individuals who either resist the innovation or have not adopted it due to uncertainty (Rogers 2003:281).

The original model, and most of the research, relates to the diffusion of innovations to *individuals* in open social systems, where each individual usually goes through the classic stages of the innovation-decision, which are:

1. The person learns about the new innovation (knowledge);
2. Forms an opinion about it (persuasion);

3. A decision is made to adopt or not;
4. The innovation is implemented (if decision was to adopt);
5. Confirmation stage, where the person reflects on his/her decision and either confirms it was the right one or now rejects the innovation (Rogers 2003:28).

However, in this study, the discussion relates to the diffusion of innovations through an *organisation* (the CoP), in which case there is a difference in the innovation-decision dynamics. In such settings, an *authority* innovation-decision is made (to adopt or not) after discussion by the leaders of the organisation (Rogers 2003:28). The rest of the organisation is then forced to accept and implement the innovation. Therefore, in such cases the individual in the organisation might first learn that a decision has been made *and* implemented (no.3-4 above), before learning what the innovation is about (no.1), then form an opinion (no.2) and subsequently confirmation or rejection takes place (no.5).

In the context of this study, the leaders of the organisation (CoP council) made an *authority* innovation-decision to introduce an innovation (standard setting) into the core-business of the organisation (assessment). Subsequently, the organisational members (CoP examiners) were exposed to a change process. The diffusion of this innovation was, therefore, measured in how many examiners support the innovation's application in the CoP. Adoption was defined as supporting the innovation and non-adoption represented resistance or uncertainty regarding the innovation.

The CoP council took the decision in May 2011 (Time 0) to gradually implement the Cohen method over time as discussed in Chapter 1. However, the diffusion of the standard setting in the 'hearts and minds' of the examiners was a critical factor for the long-term sustainability or refreezing capacity of the innovation. This was the motivation to study the diffusion of standard setting through the most important stakeholder, the examiners at the time.

Table 2.6 contains the essential components of the diffusion of innovation definition and contextualises each component to the corresponding process of change and innovation as they took place in the CoP during the study period.

TABLE 2.6: THE TRANSLATION OF THE DIFFUSION OF INNOVATION DEFINITION TO THE PRESENT STUDY

Component of the definition*	Relevant component in the present study
The **Innovation**	The introduction and implementation of standard setting (The Cohen method)
The **Communication channels**	The published FCP regulations on the CMSA website, the educational seminar for the CoP examiners in 2013, FCP council meetings since May 2011 and by word of mouth
The **Time** frame	**Time 0** – May 2011 (standard setting presented to the CoP council by the champion – Prof Vanessa Burch) **Time 1** – Feb 2013 (Round 1 of survey) **Time 2** – Feb 2014 (Round 2 of survey)
The members of the **social system**	The 54 current CoP examiners

* Rogers 2003:11

Summary

Schein concluded that change is probably better defined as a "leaning" process and the planned change, and change management relates closely to a form of "managed learning" (Schein 1996:7). Change targets need to unlearn old attitudes and behaviours and learn new ones. The discomfort and psychological anxiety with this cognitive restructuring process needs to be minimised by change agents and champions by providing the appropriate psychological safety measures.

2.4 CONCLUSION

The nature of standard setting or assessment calibration in medical education, irrespective of the particular methodology of any individual method, is fundamentally a judgement decision by educators, who are responsible for making pass/fail decisions about examinees (Cizek *et al.* 2004:33). The debate in the general education and medical education literature has moved on from 'if standard setting should be used' to 'which methods are the most appropriate for a given context'.

This chapter has reviewed and discussed the relevant literature to conceptualise and contextualise the topic of standard setting within the broader domain of assessment in medical education. The specific context in this thesis is the introduction of standard setting for the *written* components of the postgraduate medical licensing examinations, conducted by the CoP, for specialist physician trainees in South Africa.

Several important issues relevant to the topic were discussed and are summarised below.

High quality standard setting cannot take place in the absence of high quality assessment data. Good decisions about candidate progression cannot be made using poor quality test data.

The Angoff standard setting method and many of its common modifications, as well as the relatively new Cohen method, were reviewed and discussed. Comparisons between these two methods, and other methods described in the literature, were discussed.

Medical educators are accountable to a range of stakeholders (trainees, employers, the public) to develop and use high quality assessment and standard setting strategies.

The theory of change and change management, with particular reference to initiating and managing a change process in medical education and the accompanying diffusion of a new innovation in an organisation, were explained and discussed using the relevant literature.

The utility parameters of a standard setting method were derived from the literature and formulated into a user-friendly framework, which was eventually used to evaluate the Angoff and Cohen methods, using data derived from the experimental components of this study (Chapters 4-8).

In the next chapter, Chapter 3, the research design and methodology used in this study to address the research problems and research questions posed in Chapter 1, are described and discussed.

CHAPTER 3

RESEARCH DESIGN AND METHODOLOGY

3.1 INTRODUCTION

This chapter deals with the research design and methodology employed to address the research problems and research questions posed in Chapter 1. This research study had two components, each with its own objectives and methodology. They were:

- Evaluation of the impact of introducing standard setting in the CoP; and
- Comparison of the performance and overall utility of two standard setting methods as used and evaluated in this study.

This chapter provides clarity and an understanding of the methods used in each of the two research components. To facilitate orientation between the research questions, the related objectives and the methods employed to answer the questions and satisfy the objectives, the research questions and the objectives are restated in each section, followed by a detailed description of the methods used. In addition, some literature supporting the methods used in this study, are also discussed.

The data analysis from both the first and second components of this study was done by the principal researcher, in conjunction with the study promoters, a private research and statistics consulting company, MelodyM Consulting, and the Department of Biostatistics of the University of the Free State.

The chapter commences with a description of the research paradigm and design, followed by an explanation of the methods used for the different sections of the study. The first section is the literature review, followed by the respective sections on research components 1 and 2. The chapter ends with a brief overview of the key issues of the chapter.

3.1.1 Research paradigm of this study

A research paradigm is the philosophical departure point or worldview of a researcher as it relates to his/her methodological approach and research design (Tavakol &

Sandars 2014a:747; Trafford & Leshem 2008:94-97). In this study, a *positivistic* paradigm (Tavakol & Sandars 2014a:747) was used and followed. That implies that a predominantly quantitative, deductive approach to theory and hypothesis testing was employed in this study (Trafford & Leshem 2008:97), with the aim of generating new knowledge (Tavakol & Sandars 2014a:748,749) in the field of standard setting.

3.2 RESEARCH DESIGN

Action research

As briefly mentioned in Chapter 1, the research design used for this study was an *action research* approach. This formed the theoretical framework of the study (Tavakol & Sandars 2014a:752).

Kurt Lewin founded and described action research for the first time in his 1946 paper entitled "Action research and Minority Problems" (Lewin 1946). He discussed this new strategy to study a change process; in the case of his 1946 paper it pertained to social change. Lewin also proposed that action research should be the preferred approach to study a change management process (*cf.* Riel 2010), as was done in this study.

Riel (2010) provides a pragmatic definition of action research as "Action research is an experiment in design, and involves implementing an action to study its consequences". McNiff and Whitehead (2010:5) further explain that action research is about taking action to improve practice and then going further to research the impact and effects of the implemented action, to gain new knowledge and insight on the changed practice. This was the research framework adopted for this study, which aimed to improve assessment practice in the CoP by introducing standard setting (the action), while simultaneously researching the impact and effects over the 30 month change management process (the research aspect).

In action research strategy is regarded by many authors as a powerful form of educational research (McNiff 2013:24), since it "involves learning in and through action and reflection, and is conducted in a variety of contexts", including health care (McNiff 2013:24). McNiff explained that there is no such 'thing' as action research per se, it is merely a strategy to provide a theoretical framework to explain the research approach undertaken in a study (McNiff 2013:24).

The goals or aims of action research according to Riel (2010) include:

- Improvement of professional practice through problem solving and continued learning;
- Developing a deep understanding of practice and the specific theory driving the change actions implemented; and
- Improvement in the community where the practice is embedded by participatory research efforts.

The aim of the present study was well aligned with these goals, which provided an ideal framework in which to embed the respective research methods used for the two components of this study.

The next section describes the strategy used for the literature review. Thereafter follows a description of the individual methods used in each of the two research components of the study. Apart from the literature review, the research methods used to investigate the two components of this study were predominantly quantitative in nature.

3.3 LITERATURE REVIEW

Research Question 1: How is standard setting of assessment processes in medical education conceptualised in the literature and contextualised in postgraduate specialist certification examinations offered by the Colleges of Medicine of South Africa?

Related objective 1.4.3.1: Conceptualise the role of standard setting as it pertains to assessment in medical education and contextualise it to postgraduate written assessments for specialist physicians in South Africa.

In Chapter 2, a comprehensive discussion and synthesis of the relevant literature was provided to enable the conceptualisation and contextualisation of standard setting in medical education. A comprehensive literature search was conducted using PubMed®, ERIC® and Google Scholar® as the main search engines. The key search terms in titles and abstracts were: "standard setting", "cut-score", "cut score", "written test*", "written exam*", "written assess*", "assess*", "assessment calibration" "pass standard", "passing score" and "pass mark" as well as "Angoff", "Cohen" and "Cohen-

Schotanus". In addition, the literature regarding social accountability, change management and diffusion of innovations was also reviewed and integrated into the review in Chapter 2.

Document review

The South African context about standard setting in medical education, which falls under higher education and training, was sourced from current policy documents on assessment in higher education by the South African Qualifications Authority (SAQA), the health professions education regulator in South Africa, the Health Professions Council of South Africa (HPCSA), the CMSA and the CoP.

3.4 EMPIRICAL RESEARCH COMPONENT 1 – QUESTIONNAIRE SURVEY AND SEMINAR

Research Question 2: What are the knowledge, attitudes, views and perspectives of CoP examiners regarding standard setting, and do they change with training and exposure to a process of standard setting?

Related objective 1.4.3.2: Determine the knowledge, attitudes, views and perspectives of the CoP examiners about standard setting.

Related objective 1.4.3.3: Design, deliver and evaluate the impact of a seminar dealing with standard setting in the CoP.

Research Question 2 and objectives 1.4.3.2 to 1.4.3.3 related to the CoP examiners' knowledge, attitudes, views and perspectives about standard setting. For this component, a prospective cohort study design was used to answer and address this research question and the two related objectives. This part of the research study also contributed to answering Research Question 3 and the final related objective (1.4.3.7), as stated below.

Related objective 1.4.3.7: Use the findings of Objectives 1.4.3.2 – 1.4.3.6 to contribute to the evaluation of the utility (as defined in Chapter 2) of the Cohen method, as compared to the Angoff method, for the written FCP (SA) examinations.

A prospective cohort design involves identifying a study population and observing or measuring what changes occur within that population after exposure to an intervention (Morroni & Myer 2007:81). In this component of the study, the population under investigation was the examiners of the CoP. The interventions they were exposed to (the standard setting processes and educational seminar), as well as the instruments used to measure or evaluate their impact (the online questionnaire survey and seminar rating sheet), are described below.

3.4.1 An online questionnaire survey

3.4.1.1 *Development of the questionnaire survey*

In order to conduct an initial situational analysis at the start of the study on the knowledge, attitudes, views and perspectives of the CoP examiners about standard setting, the most cost-effective and inclusive methodology available to the researcher was a questionnaire survey. Due to the fact that the examiners involved in the Part I (entry) and Part II (exit-level) of the Fellowship of the College of Physicians (FCP) of South Africa (SA) examinations are located all over South Africa at its eight medical schools, the decision was taken to opt for an online, web-based questionnaire survey. The link to the questionnaire survey was emailed to each questionnaire participant (FCP examiner) who could then click on it to access the questionnaire.

The survey questions were developed by consulting the available literature on academic staff members' knowledge, attitudes, views and perspectives about standard setting in general, as well as by gathering additional inputs from informal discussions with the current (2010-2014) council members of the CoP, all of whom are academic staff members at their respective medical schools and examiners for the CoP.

The questionnaire comprised of a combination of single most-applicable options, multiple choice questions (MCQs) and a number of statements for evaluation using a five-point Likert-scale (Strongly agree, Agree, Uncertain, Disagree, Strongly Disagree), aimed at determining the knowledge, attitudes, views and perspectives of the examiners of the CoP about standard setting. Spaces were provided after each MCQ or statement in the questionnaire where additional, optional free text comments could be added by respondents to provide additional insight into their responses.

The questionnaire was *piloted* using a group of five specialist physicians employed at the University of the Free State (for practical reasons) who had insight into the CoP examination processes or FCP (SA) examination (current or previous CoP examiners or recent FCP graduates). This quality assurance step was needed to test the online questionnaire's clarity and enabled further refinement of the survey before sending it to the study population (Katzenellenbogen & Joubert 2007:116).

After consenting to take part in the pilot process and a briefing by the researcher, the pilot study evaluated the questionnaire with regard to clarity (to ensure unambiguousness), time taken to complete the survey, and the examiners' opinion about the ability of the survey to adequately capture their knowledge, attitudes, views and perspectives about standard setting. The feedback and inputs obtained from the pilot participants as well as a discussion with the promoters of the study were used to make minor modifications to some of the items, to produce the final questionnaire (*cf.* the full questionnaire in Appendix A-1). Survey data collected during the pilot process were not included in the study, but some pilot participants were included in the formal survey since they were part of the target population, the CoP examiners. The average time taken to complete the 16 questionnaire items was nine minutes. Participants unanimously agreed that the quantitative questionnaire items were sufficiently succinct and focused to capture their knowledge, attitudes, views and perspectives about standard setting. Although the qualitative free text comments box after each item was welcomed, as an option to potentially clarify or justify a response, the pilot participants felt it was not needed to capture their opinions. The researcher decided to keep it in the survey.

Apart from the consent section, the questionnaire survey consisted of four sections, each addressing a separate issue of interest. Table 3.1 below provides a summary of the final questionnaire's breakdown, expressed per section.

TABLE 3.1: BREAKDOWN OF SECTIONS IN THE QUESTIONNAIRE SURVEY

Section	Sub-section(s)	Item no. in survey	Total items (16)
Consent	-	1	1
1. Knowledge about standard setting	general - regarding the concept	2	1
	the Cohen method	6	1
2. Education/training on standard setting	-	3	1
3. Awareness of implementation of standard setting in CoP	-	4; 5	2
4. Attitudes, views and **perceptions** of standard setting about	changing previous fixed 50% pass mark	8; 9; 14	3
	current use of the Cohen method	10; 11; 12; 13	4
	expanded use of the Cohen method	7; 15; 16	3

The complete questionnaires, as used in round 1 and round 2 of the survey respectively, are attached in Appendix A-1 and Appendix A-2. They were identical, except that the round 2 survey had one additional item at the end about the feasibility and sustainability of the Angoff method, as it was used in this study.

3.4.1.2 *Target population*

There were 54 examiners and moderators involved in the FCP (SA) Part I and II examinations between 2010-2012. This cohort of examiners was deemed the 'current' examiners since they were involved in the FCP (SA) examinations at the time when standard setting was introduced in the CoP. They were regarded as the most influential stakeholders and 'gatekeepers' relating to the introduction and implementation of standard setting in the CoP. For these reasons, the 2010-2012 cohort of examiners (n=54) constituted the *target population* (Joubert & Katzenellenbogen 2007:94; Tavakol & Sandars 2014b:840) of this component of the study.

Qualifications and characteristics of CoP examiners

CoP examiners are all HPCSA registered, practicing specialist physicians themselves, having qualified before 2011 and hence through one or both of the following specialist examination routes - FCP (SA) or M.Med. They work in academic medical centres in the public sector of South Africa and have considerable experience in training and examining undergraduate medical students and postgraduate specialist trainees in Internal Medicine. The vast majority of CoP examiners have additional training in a sub-specialty discipline of Internal Medicine, such as Pulmonology or Geriatrics.

How does the CMSA appoint CoP examiners?

Each year the CoP council invites potential new examiners to attend an FCP (SA) clinical examination sitting as observers. Subsequently, the CoP Council nominate appropriate examiner candidates to the Examinations and Credentials Committee (ECC) of the CMSA for ratification and final approval. The ECC is an appointed committee of the Senate of the CMSA. The Senate is the highest decision making group in the CMSA.

3.4.1.3 *Administration of survey*

As mentioned in the preceding section, an online (internet-based) questionnaire survey was employed in this component of the study. This survey approach, supported by the literature (Ritter, Lorig, Laurent & Matthews 2004), offered considerable advantages in terms of cost and time effectiveness in the data gathering process (Tavakol & Sandars 2014b:843). SurveyMonkey® (SurveyMonkey Inc. 2013) was used to administer and gather the responses of the online survey.

An *online web-link* to the survey was emailed to *all* of these 54 'current' examiners by the Secretary of the CoP, explaining the importance of their involvement and inviting them to participate in the research study.

The *first round* of the survey was administered in February 2013 (Time 1, *cf.* Figure 1.3), 18 months after introducing standard setting in the CoP. Once respondents received and opened the emailed survey link, they were prompted to read the consent form and tick a box indicating that they were consenting to participation in the survey and that their response data could be used in the study. The survey was programmed in such a manner that unless the respondent consented, he or she could not continue with the survey process. Since the target population was quite small (n=54), a good

response rate (>70%) was needed to substantiate the findings from this component of the study (Tavakol & Sandars 2014b:843) with a less than 10% margin of error as calculated by the survey software used (SurveyMonkey Inc. 2014). Three follow-up emails were sent, at weekly intervals, to maximise participation in the survey.

The data obtained from the first round of the survey, the initial situational analysis in February 2013 (Time 1), formed the basis for the development of a customised 60-minute academic staff development *seminar* on standard setting for the CoP examiners. The seminar specifically addressed issues noted in the survey data. This seminar was conducted on two occasions during 2013, one in the south-west (Cape Town) and one in the north-east (Durban) of the country.

The *second and final round* of the survey was sent to the same cohort of CoP examiners (n=54) 12 months later in February 2014 (Time 2, *cf.* Figure 1.3). A similar weekly email strategy was again followed to maximise participation. The same questionnaire was used in the second round of the survey, with one additional question added for respondents who participated in any of the five Angoff standard setting meetings, conducted as part of this research project. This additional question was aimed at measuring the examiners' views as to the long term feasibility and sustainability of the Angoff standard setting method for the written components of the FCP (SA) examination. Five Angoff standard setting meetings took place between October 2011 and October 2013 during the CoP council and examiners meeting at the start of each of the four-day FCP (SA) clinical examination events, which are conducted biannually in May and October.

This second situational analysis was conducted after the examiners had 30 months of exposure to standard setting processes in the CoP and the educational seminar. The aim of the second analysis was to determine if there had been any changes in their knowledge, attitudes, views and perspectives about standard setting, *as a cohort* of examiners, since the initial survey.

Potential changes over time in the responses of CoP examiners in the different sections of the survey (*cf.* Table 3.1) were important. For this reason, individual responses from the items of the survey were paired for data analysis for those examiners who participated in both rounds of the survey. This was needed to enable appropriate statistical analysis of the survey data. The coding was done by an external person, not

involved in the study or CoP, to ensure the respondents remained anonymous to the researchers at all times.

3.4.1.4 *Data analysis of the survey*

The analysis of the survey data is described in relation to Table 3.1, which outlines the different sections of the survey. The analysis for **Sections 1-3** (the Knowledge, Education/Training and Awareness sections), comprising five survey items, related to:

- Determining the *initial* position (in February 2013, Time 1) of the CoP examiners *as a cohort*. This was after 18 months of exposure to standard setting in the CoP.
- Determining the *current* position (in February 2014, Time 2) of CoP examiners *as a cohort*. This was after 30 months of exposure to standard setting in the CoP and the educational seminar on standard setting.
- Determining if there were *any changes* in the CoP examiners between round 1 and round 2 of the survey *as a cohort*, as well as on an *individual* level for those examiners who completed both rounds of the survey (the matched data).

Section 4 (*cf.* Table 3.1) of the survey, was a critical part of the survey which measured the 'hearts and minds' of the CoP examiners over the course of the change management process in the CoP regarding standard setting. The section consisted of the last ten items in the survey, which evaluated the attitudes, views and perspectives of the CoP examiners about standard setting.

Section 4 had three sub-sections which explored the attitudes, views and perspectives of the CoP examiners on three levels of change relating to the introduction and implementation of standard setting in the CoP. These were pertinent at the time of administering round 1 of the survey in February 2013. They were:

1. Changing the traditional past practice of using a fixed 50% pass mark for all the written papers in the FCP (SA) examination.
2. *Current use* of the Cohen method in two written components of the FCP (SA) examination – the Part I MCQ and Part II Objective Test papers.
3. *Potential expanded use* of the Cohen method to include the final written component in the FCP (SA) examination, the Part II SEQ test.

The analysis of these three sub-sections of Section 4 was done in the same way as for Sections 1-3, which was described above.

Subsequently, based on the change management and diffusions of innovation literature, which were reviewed and discussed in Chapter 2 (section 2.3.7), the survey participants were classified in one of two broad categories, based on their responses to Section 4 of the survey. They were:

- *Adopters* - They were supporters of changes made and further changes proposed; or
- *Non-adopters* - They were either uncertain or did not agree with the changes made and further changes proposed

Statistical analysis was done using Microsoft Excel® (Microsoft 2007) and SPSS® Inc. (IBM Corp 2013) software. The free text comments were read, analysed and summarised to provide some further insight into the responses from the CoP examiners. The tests used to determine if there were any statistically significant changes in views, using matched data, were the Marginal homogeneity tests (Mehta & Patel 1989:73) (as an omnibus test), followed by post-hoc McNemar tests (Altman 1991:259; Mehta & Patel 1989:70), with the appropriate Bonferroni adjustment (Altman 1991:221; Weisstein 2014) if applicable. The results from this component of the study are provided, explained and discussed in detail in Chapter 4.

3.4.2 The educational seminar on standard setting

3.4.2.1 *Content of the seminar*

The situational analysis data from the initial survey formed the basis of the customised educational seminar presented to CoP examiners. The concept, principles and importance of standard setting in assessment were explained and discussed. The two specific methods investigated in this study, the Angoff and Cohen methods, were explained and discussed in detail, highlighting their respective advantages and disadvantages. Examiners were reminded/informed that the Cohen method had been selected and introduced by the CoP council in 2011. They were also informed that the performance of the Cohen method was being reviewed and compared to the Angoff

method, as a possible alternative, under formal research conditions as part of this study. The full Microsoft PowerPoint® presentation is attached in Appendix B-1.

3.4.2.2 *Evaluation of the seminar*

The analysis of the evaluation of the standard setting seminar was focussed on determining the pre- and post-seminar understanding and opinions of the CoP examiners regarding a number of *utility parameters* of the Angoff and Cohen methods of standard setting.

The CoP examiners attending the seminar were provided with the same rating sheet before and after the seminar to note their responses. The seminar evaluation form is attached at the end of the thesis as Appendix B-2. No names of examiners were requested, hence their responses were anonymous. The sheets were numbered per examiner and their pre- and post-seminar responses were paired to enable statistical analysis. The data gathered from the seminar's evaluation were used, in conjunction with the rest of the study's data, to inform the *utility evaluation* of the Angoff and Cohen methods. This contributed to answering Research Question 3 (*cf.* section 1.3.4) and the related objective (*cf.* 1.4.3.7), which are provided again at the start of the next section.

The rating sheet used to evaluate the seminar's impact was developed to capture the understanding and opinions of the seminar participants (CoP examiners) on seven utility parameters of the Angoff and Cohen methods respectively. They were: Objectivity, Feasibility, Sustainability, Credibility, Validity, Reliability and Transparency. These parameters were used because they evaluated the examiners' perspectives on the inherent *methodology* of the two standard setting methods, as well as the cognitive and procedural *functioning* of methods, as used and compared in the study.

Participants were asked to rate the Angoff method and Cohen method (as used in this study) for each parameter, before and after the seminar. Each parameter had a short description/definition included on the sheet, to aid the participants in their conceptual understanding and judgement of each parameter, for each method. Each parameter was rated on a 7-point ordinal rating scale, with an additional option of "I don't know" provided. This approach is supported by literature on pre/post-test surveys (Spears & Wilson 2010:2).

3.4.2.3 *Data analysis of the seminar*

The data were analysed using Microsoft Excel® (Microsoft 2007) and SPSS® Inc. (IBM Corp 2013) for descriptive statistics and the Wilcoxon signed-rank test (Altman 1991:203; Statistics Solutions 2014), to determine if there were any statistically significant differences between the non-parametric ordinal ratings provided by the examiners. These results are provided, explained and discussed in detail in Chapter 4, with an overarching discussion, in conjunction with the second component of the study, in Chapter 8.

3.5 EMPIRICAL RESEARCH COMPONENT 2 – COMPARITIVE STUDY OF TWO METHODS

Research Question 3: Is using the Cohen method, as compared to Angoff method, a more appropriate way of determining the pass mark for the written assessments of the FCP (SA) examination (Part I MCQ test, Part II Objective test and Part II SEQ test)?

Related objective 1.4.3.4: Determine the performance of the Angoff method of standard setting using five cycles of written FCP (SA) examinations data.

Related objective 1.4.3.5: Determine the performance of the Cohen method of standard setting using the same five cycles of written FCP (SA) examinations data as in 1.4.3.4.

Related objective 1.4.3.6: Determine the variability of the scores of the top-performing candidates, in order to substantiate the assumption within the Cohen method that the test scores of the top candidates sitting the FCP (SA) written examinations, are stable and, therefore, comparable.

Related objective 1.4.3.7: Use the findings of Objectives 1.4.3.2 – 1.4.3.6 to contribute to the evaluation of the utility (as defined in Chapter 2) of the Cohen method, as compared to the Angoff method, for the written FCP (SA) examinations.

Research Question 3 and objectives 1.4.3.4 to 1.4.3.7 relate to the evaluation of the utility of the Cohen method of standard setting for the written assessments of the FCP (SA) examination. For this component, a comparison of the two standard setting methods (Cohen and Angoff) was done using a comparative study design (Altman 1991:6), which entailed analysing the *same* five cycles of the written FCP (SA) examinations, using two different methods, and comparing their results. The specific methodologies used during this component of the study are discussed below.

3.5.1 The Angoff method of standard setting

The *original* Angoff method described by William Angoff (Angoff 1971:515) is commonly referred to as the *Yes/No* Angoff method (Downing *et al.* 2006:53). This version of the Angoff method, as well as the common modifications and their challenges were discussed in detail in Chapter 2 in the context of the relevant literature.

3.5.1.1 *Angoff method in the context of the CoP examiners' meetings*

For the purpose of the study it was imperative that the Angoff method be conducted to the highest possible standards, in the context of the relevant resource limitations of the CoP. The reason for this approach was not only to produce useful Angoff results for the study itself, but also to determine whether use of the method could be sustained beyond the data collection period as a possible alternative to the Cohen method, if the Angoff method's utility was deemed favourable by the findings of this study.

Angoff version used

Owing to the *limited experience* of CoP examiners, regarding standard setting procedures, and the *time pressures* due to the large number of agenda items discussed at the biannual Council meetings of the CoP examiners during the Part II FCP (SA) clinical examinations, the researcher and his promoters opted for the original (Yes/No) version of the Angoff method to set the pass marks for the written papers of the FCP (SA) examination included in the study. Suggestions from the literature indicated that this version of the Angoff method was easier to explain, comprehend and execute in the context of novice Angoff judges (Chinn & Hertz 2002:7; Impara & Plake 1997:355).

Another motivation for using the Yes/No Angoff method in this study, as explained in section 2.3.1 in Chapter 2, was that it is also used for the written certification examinations of one of the largest and most established certification bodies in the Internal Medicine globally, the American Board of Internal Medicine (ABIM) (Tormey 2014:8).

Normative data and iterative discussion rounds with re-rating

At each biannual CoP council meeting the next cycle's written papers are reviewed. To align with this review practice, the different test papers included in this study were, therefore, submitted to an Angoff process *prospectively*. The only exception to this was the SEQ papers, which were retrospectively submitted to the Angoff process, due to the timing of the commencement of the study and the possible change to future SEQ tests' formats (*cf.* further explanation below). Since each paper of the three written test formats are unique for every cycle, with only the MCQ tracker items retained for research purposes, there were no item performance data available to use as part of a modified-Angoff procedure – iterative discussion rounds and subsequent review and re-ratings of items by the judging panel.

Concerns have been expressed in the literature (*cf.* Chapter 2, section 2.2.9.2) that the lack of normative test item data, during iterative discussions and review rounds, may result in a process where the panellists converge in the *wrong* direction during discussions and subsequent re-ratings, thereby *reducing* the accuracy of item difficulty probabilities. Due to this concern, coupled with the time constraints during the CoP Council meetings, the decision was taken to *not* use iterative discussion rounds and subsequent review and re-ratings during this study's Angoff processes. Therefore, a pure absolute/criterion-referenced, panel-based, test-centred approach was followed in this study.

Angoff panellist selection

Each year a cohort of examiners is selected by the CoP council to conduct the biannual FCP (SA) clinical examination, from the pool of 54 approved CoP examiners. These cohorts were invited to become part of the Angoff panels at each meeting. Therefore, while the Angoff panels were a form of convenience sampling, there were no known biases involved in the cohort selections. There were 11-18 examiners present at each

meeting, which represented a sufficient number of participants for an Angoff process (Brandon 2004:68; Downing *et al.* 2006:51; Norcini 2003:467).

Therefore, for the Part I MCQ paper, the above Yes/No method was used and for the constructed-response papers (the Part II Objective Test and SEQ papers) the Yes/No Angoff method was modified to ask the judges what a borderline candidate would score on each item in the paper. Section 3.5.1.2 below describes the exact questions used in the Angoff processes.

3.5.1.2 *Yes/No Angoff method procedure in the CoP examiners' meetings*

A short information and invitation letter was provided to the examiners at the start of the Angoff meetings, inviting them to participate in the Angoff process. Examiners, who were willing to participate in the research project, were requested to complete the consent form which was attached at the end of the letter. The letter and consent form are attached in Appendix C.

The process which was followed during the Yes/No Angoff meetings was as follows:

1. A careful explanation regarding the Yes/No Angoff method procedure and what is expected of each panellist.
2. A thorough discussion of what constituted a 'borderline' or just-competent candidate for the respective entry- and exit-level papers, with the aim of reaching a consensus perspective.
3. The panellists then proceeded to judge each item of the respective papers *individually*.
4. The questions which were put to each panel member to judge each item of the three papers were:

    a. <u>Part I entry MCQ test – 150-item single best answer MCQ paper</u>:
    Rate each MCQ item: "Would a BORDERLINE candidate answer this question CORRECTLY?" YES or NO (scored 1/0 respectively for all 150 items). Model answers were provided.

b. Part II exit Objective Test – 30-item constructed-response paper:
Rate each 7-mark item: "What would a BORDERLINE candidate score out of 7 for this question?" Holistic item judgement, out of 7 marks, for all 30 items. Model answers were provided.

c. Part II exit SEQ test – 20-item constructed-response paper:
Rate each 15-mark item: "What would a BORDERLINE candidate score out of 15 for this question?" Holistic item judgement, out of 15 marks, for all 20 items. Model answers were *not available* and hence not provided.

5. As mentioned before, no 'reality checks' with further iterative judging rounds, subsequent to a discussion on the items, were held. After the single round of judging the difficulty of the items for a borderline candidate sitting the paper, the anonymous judgement sheets were collated and returned to the researcher for data input and analysis. The only other information gathered from the judges were the amount of experience, in years, they had as a CoP examiner.

The scores of all the judges were first summated and averaged across each individual question item in each respective paper and then, finally, all the items within each paper were averaged to derive an Angoff pass mark for each of the three papers. In addition, the Angoff pass marks for each paper, as judged per individual panel member, were also calculated by averaging the item scores for each panellist for each paper.

3.5.2 The Cohen method of standard setting

The Cohen method was explained and discussed in Chapter 1 (Section 1.2.6) and Chapter 2 (Section 2.2.10). It was described and published for the first time in Dutch in 1996 (Cohen-Schotanus *et al.* 1996) and then again in English in 2010 (Cohen-Schotanus & Van der Vleuten 2010:157).

The CoP applies the **Cohen65 model** of the Cohen method to set pass standards for their written tests. The CoP expects candidates to have a minimum of 65% knowledge of the test material, after correction for difficulty, as assessed by the different written FCP (SA) assessments. The 95th percentile mark of the test results, for a particular test, is multiplied by 65% to derive the Cohen65 pass standard (pass mark).

To address the research question and related objectives relevant to this component of the study, the performance and appropriateness of the Cohen method was compared with the Angoff method, as explained in section 3.5.1 above. The same five sets of FCP (SA) examination data, as described in Table 3.2, were used to derive a pass mark, and resultant failure rate, using both the Cohen method and the Angoff method. Thereafter, a comparison of these data was done using the appropriate statistical tests.

3.5.2.1 *The stability of the top performing candidates*

In addition, the fundamental assumption upon which the Cohen method is based was also evaluated, i.e. that the marks of cohorts of top-performing students are relatively stable from one assessment to the next. To test this assumption the stability of the 95th percentile mark, i.e., the range between top cohorts sitting the same component of the FCP (SA) examinations (Part I MCQ test or Part II OT or SEQ test) was evaluated. The 95th percentile was used because the paper by Cohen-Schotanus and Van der Vleuten (2010:159) reported that this percentile offered the best balance between performance stability and reliability of the top performing candidates.

To allow for meaningful comparisons of the performance of the top-performing candidates in consecutive cohorts, they needed to be exposed to the same test items across the five examination cycles. Due to concerns about the security and integrity of the individual examination papers, this could only be done for the Part I MCQ papers. Thirty 'tracker' items were selected from 150 test items in the Part I MCQ items (20% of a paper's items) were selected across a wide blueprinted range of items and included in each of the five cycles of the Part I MCQ paper included in the study. This equating step enabled a direct comparison of the candidates' performance across the five cycles of the Part I MCQ test. The same marking criteria were applied to the tracker items, as to the overall MCQ paper.

3.5.3 Data included in the study

Five cycles of FCP (SA) Part I and Part II written examinations data, collected from August 2011 to February 2014, were included and analysed in the study. See Table 3.2 for the details.

Data collection for the five Part I MCQ tests and Part II OT cycles starting with the first exam sitting in 2012 and ended with the February 2014 sitting. Data collection for the Part II SEQ test started one sitting earlier, in August 2011, because there was a possibility that the SEQ test format was due to change in 2014.

The Secretary of the CoP provided the examination papers to the panel of examiners at each examiner meeting during the biannual FCP (SA) Part II clinical examination events.

TABLE 3.2: FCP (SA) WRITTEN EXAMINATIONS INCLUDED IN THE STUDY

Examination name	Format/Type	Exam cycles
Part I MCQ test (entry exam)	**150 MCQ items** selected-response items, single best answer from 5 possible options (with 30 tracker items repeated in all 5 cycles)	Mar 2012
		Aug 2012
		Jan 2013
		Jun 2013
		Feb 2014
Part II Objective test (OT) (exit exam)	**30x 7-mark items** constructed-response **clinically focused** short answer question items	Mar 2012
		Aug 2012
		Feb 2013
		Jul 2013
		Feb 2014
Part II Short Essay Question (SEQ) test (exit exam)	**20x 15-mark SEQ items** constructed-response **theory** short essay questions	Aug 2011
		Mar 2012
		Aug 2012
		Feb 2013
		Jul 2013

3.5.4 The Candidates

The cohorts of candidates sitting the *entry-level* Part I FCP (SA) MCQ test are unselected and hence, per definition, they are a *heterogeneous* group in terms of examination preparation, number of attempts at this examination, academic ability, demographic composition and work location in the country. The only common feature they have is their desire to become specialist physicians. The specific qualification criteria to allow a candidate to sit the examination are described in Chapter 1. The examination can be undertaken at multiple examination centres across South Africa.

Candidates sitting the *exit-level* Part II FCP (SA) Objective Test (OT) and SEQ test are more *homogeneous*; they have all passed the Part I MCQ test and have already completed or are nearing completion of their specialist (residency) training to become specialist physicians. Given these stringent entry requirements, the Part II cohorts are homogeneous in nature in terms of examination preparedness. The specific qualification criteria to allow a candidate to sit the Part II exit examination were described in Chapter 1 (section 1.2.2). The examination can be undertaken at multiple examination centres across South Africa.

3.5.5 Marking systems used in the CoP for each test format

As per the regulations governing the administration of the 150 item *FCP (SA) Part I MCQ* test, it is subject to negative marking to discourage guessing in the test. For every correct answer, 1 mark is allocated, 0 = not answered and -0.25 for an incorrect answer. The test answer-sheets from across the country are sent to the CMSA central examinations office in Johannesburg, South Africa, were they are scanned and electronically marked. The results are then sent to the convenor of the MCQ test, who checks the results, after which they are published by the central CMSA office in Johannesburg.

For the purpose of the study (2012 - 2014), the raw results were sent to the researcher who determined the Cohen65 pass mark. The Cohen65 pass mark was then applied to the results, after which they were reviewed and approved by the convenor, prior to publication by the central CMSA office in Johannesburg.

The *FCP (SA) Part II OT* items are 7-mark, constructed-response short answer (words or short phrases) items, which are marked by hand using a model answer. The candidates' answer-booklets from across the country are sent to the CMSA central examinations office in Johannesburg, South Africa, were they are collated and sorted into item order. Six markers then each receive a batch of answer-booklets to mark. Each examiner marks five questions and submits the marks to the convenor who collates and reviews the marks prior to publication by the CMSA office in Johannesburg. The original marked answer-booklets are sent back from the markers to the CMSA central examinations office in Johannesburg for archiving.

For the purpose of the study (2012 - 2014), the raw results were sent to the researcher who determined the pass mark using the Cohen65 method. The Cohen65 pass mark was then applied to the results, after which they were reviewed and approved by the convenor, prior to publication by the CMSA office in Johannesburg.

The *FCP (SA) Part II SEQ* test items are 15-mark, constructed-response, short essay items which are marked by hand. The candidates' answer-booklets from across the country are sent to the CMSA central examinations office in Johannesburg, South Africa, where they are collated and sorted into item order. Ten markers each receive a batch of answer-booklets covering the two items they submitted for the test. These are scored and the results are sent to the convenor, who collates and checks the results, before being published by the CMSA office in Johannesburg. The marked original answer-booklets are sent back from the markers to the CMSA central examinations office in Johannesburg for archiving.

For the duration of the study (2011 - 2013) the raw results were sent to the researcher who calculated the Cohen 65 pass mark. The Cohen65 pass mark was then applied to the results, after which they were reviewed and approved by the convenor, prior to publication by the CMSA office in Johannesburg. The use of model answers could not be confirmed in each test included in this study.

The OTs and SEQ tests must be passed *independently* for a candidate to gain entry to the final component, the FCP (SA) Part II clinical examination. Therefore, a *conjunctive* standard setting approach is used by the CoP (McKinley & Norcini 2014:106).

3.5.6 Data analysis on the performance of the Angoff and Cohen methods

The results for the second component of the study are provided, explained and discussed in detail in Chapter 5 (MCQ test), Chapter 6 (OT) and Chapter 7 (SEQ test), with an overarching discussion in Chapter 8.

3.5.6.1 *Descriptive statistics*

For the second component of the study, the results from the FCP (SA) written papers included in this study and the marks achieved by the top cohorts of students, were

evaluated using basic descriptive statistics (maximum score, minimum score, range, median, mean score, standard deviation and 95% confidence intervals). Microsoft Excel® (Microsoft 2007) and SPSS® Inc. (IBM Corp 2013) software were used to do the analysis in this second research component as well.

3.5.6.2 *Item analysis data*

As discussed in Chapter 2 (section 2.2.4.3), the two most important test item psychometric parameters, namely item difficulty and item discrimination index, were evaluated in this study. In addition, a new composite parameter, the item quality index, was also used and its composition is described below. A brief description of each parameter is included below for ease of reference.

Item difficulty

For the *selected-response MCQ tests*, this was established using the Proportion Correct value (*PC-value*). The PC-value of an MCQ item is determined by summing the number of candidates who answered the item *correctly* and dividing it by the *total number* of candidates who sat the examination. It is an indication of the difficulty of the item, as determined by the cohort, and is expressed as a value between 0-1. For example: If 58% of the cohort marked the item correctly, the PC-value is 0.58. The "PC-value" term was used in this thesis rather than the more traditional "p-value" (proportion-value), to avoid confusion with the inferential statistic used to indicate statistical significance.

For the *constructed-response (CR) OT and SEQ tests*, this was established using the Item Difficulty value (*ID-value*). This value is similar to the Proportion Correct or PC-value of an MCQ item, but since each OT or SEQ was scored out of seven or 15 marks respectively, and not binary (1 or 0) such as the MCQ items, the ID-value is used and calculated by *averaging* the scores of the candidates on each constructed-response item (OT or SEQ test). It provides an indication of the difficulty of the item, as determined by the cohort, and is expressed as a value between 0-1. For example: If the average cohort score was 48% for a CR item, the ID-value is 0.48 for that item.

Item Discrimination Index (DI)

The DI of an item is its ability to discriminate between the top and the bottom performing candidates on the overall test. The classical method to determine the DI

of an item is to calculate the difference between the mean performance on the item by the top and bottom 27% of the cohort in the overall test results. This method works well for a large sample size (200+ candidates), since it yields enough candidates in the respective top and bottom groups. The DI is expressed as a number between -1 and +1, where +1 means there is a 100% difference between the top group and the bottom group on the item and hence perfect discrimination. A value of -1 means that 100% of the bottom group had the answer correct and 0% of the top group, i.e. an item with flaws and an undesirable discrimination effect. An acceptable DI for an item is 0.20 or above (Downing 2009a:L4747).

Since the cohort sizes for all the tests (MCQ, OT and SEQ) did not exceed 200, it was decided to use the top and bottom 33% of candidates to determine the DIs of the test items. This ensured that the top and bottom groups of the respective cohorts had sufficient numbers of candidates to enable a sensible and more reliable calculation of the DIs.

Item Quality Index (IQI)

The IQI is the *percentage* of test items classified as 'good items'. For the purpose of the study this was defined as items having a DI of 0.20 or more *and* a PC-value % (for selected-response items) or ID-value % (for constructed-response items) of between 20% and 80%. They are, therefore, of appropriate difficulty for the group and demonstrate an acceptable level of discrimination between top and poor performing candidates. The IQI is, therefore, an indicator of the amount of good quality items contained within the test. Since the validity of the test result is influenced by the quality of the test items, as discussed in Chapter 2 (section 2.2.4.3), the IQI is important from a standard setting perspective.

3.5.6.3 *Reliability analyses of the written tests*

Both Cronbach's alpha coefficient and the Standard Error of Measurement (SEM) were calculated and reported for the tests in this study and were discussed in detail in Chapter 2 (section 2.2.4.4), with the appropriate reference to the literature.

3.5.6.4 *Reliability analyses of the standard setting procedures*

The performance of the Angoff and Cohen methods was evaluated by reviewing the *pass marks* derived by each method and the respective resultant *failure rates*.

The methodology of the Cohen method consists of applying a simple mathematical model to the performance data of the test candidates. Since this model, Cohen65 in the CoP's case, is *consistently* applied to the test data, the reliability of the Cohen method is, therefore, perfect and the coefficient alpha value is 1. This means that for a given set of test results, the Cohen method will always produce the identical pass mark, as long as the specific chosen model is consistently applied.

The reliability of the Angoff processes, however, is not perfect since it is based on item-by-item human judgement, which is known to be variable (*cf.* Chapter 2, section 2.2.9.2). In this study, the Angoff reliability was determined using three methods described in the literature. This represented an attempt to triangulate the findings. The methods used were:

1. The *standard error of the mean Angoff pass mark* of the judges, compared to the standard error of measurement of the candidates' test scores (Cohen, Kane & Crooks 1999:364). This method described by Cohen *et al.* (1999:364) determines the reliability of the Angoff process based on how it relates to the test's reliability, the Standard Error of Measurement (SEM). It has been suggested that the standard error of the mean Angoff pass mark, as generated by the panellists, should not be more than 50% of the SEM of the test.
2. The *standard deviation of the Angoff pass mark* of the judges, compared to the standard deviation of the test scores achieved by the candidates (Meskauskas 1986). This method described by Meskauskas (1986:187-203) determines the reliability of the Angoff process based on how it relates to the standard deviation of the candidates' test scores. It has been suggested that the standard deviation of the Angoff pass marks, as generated by the panellists, should not be more than 25% of the standard deviation (SD) of the candidates' test scores.
3. The Angoff *inter-rater reliability* (IRR) calculation is a measure of the internal consistency of the ratings between the panellists (George *et al.* 2006:3). The IRR for multi-rater (three or more) and binary (1/0) decisions was calculated by Light's Kappa and was used for the Angoff ratings of the MCQ tests. For multi-rater

Angoff procedures, using the scale-variable marks of the constructed-response formats [OT (0-7) and SEQ tests (0-30)], the IRR was calculated using Intra-class Correlations (ICC) (Hallgren 2012:9).

3.6 STATISTICAL SIGNIFICANCE AND ALPHA LEVEL IN THIS STUDY

The alpha level selected for use in this study was 0.05. Therefore, statistical significance was reported if *p*-values in this study were equal to or less than 0.05 (the designated alpha level). This indicated a chance of less than 5% of wrongly rejecting a null hypothesis (Howell 2007:96).

3.7 VALIDITY, RELIABILITY AND TRUSTWORTHINESS OF THE STUDY

It is important that the methods used in a study produce valid (credible) and reliable (reproducible) results and findings (Myer & Karim 2007:155). If this is not ensured, the recommendations from the study are not trustworthy and as a result hard to generalise (Norman & Eva 2010:312). In the next three sections, these aspects will be discussed as they relate to this study.

3.7.1 Validity of the study

Validity is defined as whether or not the measuring instrument actually measures what it intends to measure (Downing 2003b:830-837; Lynch, Surdyk & Eiser 2004:367; Schuwirth & Van der Vleuten 2010:196; Shumway & Harden 2003:572; Twycross 2005:36; Van der Vleuten 1996:50; 2000:1217). Therefore, in the case of a measurement of the knowledge, attitudes, views and perspectives of CoP examiners, the question was whether or not the questionnaire survey delivered and produced results that would accurately reflect the actual knowledge, attitudes, views and perspectives of CoP examiners as a whole. Although the literature could inform one on how other academic staff members around the world previously responded and reflected on similar questions and questionnaires, one cannot accurately deduce nor have a valid assessment of the knowledge, attitudes, views and perspectives of CoP examiners as a group, if one does not actually engage and ask a representative sample directly. For this reason, the entire population of CoP examiners during 2010-2012 were invited to participate in this survey.

The second matter was the quality of the actual questionnaire that was used. To ensure its validity, all the aspects that a researcher wishes to report on must be accurately evaluated within the questionnaire. Hence, careful planning of the questionnaire was critical and took place as explained in section 3.4.1 above. Probably the most important step in ensuring that the questionnaire survey was valid was the feedback from the pilot process of the questionnaire. One of the specific areas of feedback that was sought from the examiners reviewing the initial draft of the questionnaire was about whether or not they felt that the questionnaire actually measured, and would be able to reflect, their knowledge, attitudes, views and perspectives on standard setting. Feedback from the examiners who reviewed the questionnaire was reviewed and incorporated into the questionnaire, where appropriate, as explained in 3.4.1.1. The responses from the CoP examiners were anonymous and this should have promoted honest and true reflections of their knowledge, attitudes, views and perspectives on standard setting.

The validity evaluation of the second component of this study was addressed through objective 1.4.3.7. It stated that the second component of the study would determine and evaluate how the findings from the two standard setting methods compared and contrasted with one another. In other words, it would evaluate the overall utility, including validity, of these two methods to set pass standards for the different written components of the FCP (SA).

3.7.2 Reliability of the study

Reliability (or reproducibility) refers to extent to which the findings from a particular assessment can be reproduced in repeated assessments (Downing 2004:1006-1012; Lynch *et al.* 2004:367; Schuwirth & Van der Vleuten 2010:196; Twycross 2005:36; Van der Vleuten 1996:48). In the case of the questionnaire survey, the question about its reliability hinged on the extent to which the findings of the CoP examiners' responses regarding their knowledge, attitudes, views and perspectives on standard setting would be reproducible, if a different but similar (or parallel) questionnaire survey was also completed by them. Since this is not practically possible, the reliability of the findings from the actual questionnaire survey was ensured by asking multiple, but slightly different questions on the aspects under investigation in the same questionnaire (essentially creating a parallel questionnaire within the original). These questions

served as a triangulation process (Tavakol & Sandars 2014b:844) to check the reproducibility of the findings reported in the study.

The reliability of the findings of the second component of the study, a comparative study, was determined by two issues. One was the reliability of the judgements made by the Angoff panel of expert judges and the other was the analysis of the performance of the Angoff method as compared to the Cohen method of standard setting. Since the Cohen method is a mathematical model that is simply applied to the results of a written assessment it has a reliability of 1 (or 100%). The reliability of the Angoff judgements, however, is variable according to the literature (Barman 2008:959; Cusimano 1996:s116). This is mainly due to the inherent features of the Angoff method and process – human judgements on perceived minimal competence of candidates and the difficulty of a question for a borderline candidate as judged by a subject expert. This variability between judges on the same panel, one of the perceived negative characteristics of the Angoff method, was evaluated in this study. The reliability of the analysis comparing and contrasting the two standard setting methods was ensured by careful use of the appropriate statistical tests as advised by the study promoters, MelodyM Consulting and the Department of Biostatistics of the University of the Free State.

3.7.3 Trustworthiness of the study

The overall trustworthiness (or credibility) of the results of this doctoral study rested on the rationale for, and the rigour of, the methodology followed to answer the research questions and to meet the objectives of the study. In essence therefore, it refers to the *quality assurance* of the research process (Rolfe 2006:305) and its subsequent "believability" (Maykut & Morehouse 1994:64). This depends on the transparency, validity and reliability of the research (Rolfe 2006:305). In this research, transparency was ensured by openly describing the methodology used and reporting the findings that were obtained. The validity and reliability issues of this research study were discussed above.

3.8 ETHICAL CONSIDERATIONS

3.8.1 Ethics approval

Permission to conduct this Ph.D. study was obtained from the following persons or committees at the University of the Free State (UFS) (*cf.* Appendix F):

1. The Ethics Committee of the Faculty of Health Sciences – study number: 94/2012.
2. The Executive Committee of the School of Medicine (SoM), Faculty of Health Sciences .
3. The Dean, Faculty of Health Sciences.

The Vice-Rector (Academic), UFS was also informed about the study and permission to conduct the study was granted. In addition, permission to conduct the study was also obtained from the Senate of the CMSA and the Council of the CoP.

3.8.2 The first component of the study

All the CoP examiners were invited by email to respond to the online questionnaire survey. Although they were encouraged to participate, it was clearly communicated that participation in the survey was voluntary and that all responses from the online questionnaire survey would be anonymous. Every respondent had to provide consent for voluntary participation in the survey and for the data of their responses to be included in the analysis. Participants were also informed of the researcher's intent to publish the results of the survey. The cover pages of the questionnaire surveys detailed all the information just described (*cf.* Appendices A-1 and A-2). No demographic data of any participants were gathered.

3.8.3 The second component of the study

All FCP (SA) examinations data processed during the Angoff and Cohen methods were completely anonymous. The relevant examiners who were present at the examiner meetings during the biannual sitting of the FCP (SA) Part II clinical examinations were invited to participate in the Angoff procedure (*cf.* Appendix C for the information letter and consent form). All reporting of the results was done anonymously too. No demographic data were gathered from any participants in the Angoff procedures.

Since the Cohen method does not involve individuals, anonymity of this data was ensured.

3.9 CONCLUSION

This chapter has provided a discussion of the methodology and research strategies employed in the two major research components of this study. The concepts of validity, reliability and trustworthiness, as they relate to this study, were also discussed. The ethical issues relating to this study were described. In the next chapter, Chapter 4, the *results* of the first component of the study dealing with the knowledge, attitudes, views and perspectives of the CoP examiners about standard setting, is provided and explained. In addition, Chapter 4 also includes a discussion on the results of the first component of the study. Conclusions are provided in Chapter 8, after the results from the second research component have been reported and discussed in Chapters 5-7.

CHAPTER 4

INTRODUCING STANDARD SETTING IN THE COLLEGE OF PHYSICIANS OF SOUTH AFRICA – A PROCESS OF CHANGE AND DIFFUSION OF INNOVATION

4.1 INTRODUCTION

This chapter of the thesis reports on, and discusses, the findings of the first part of the study in which the knowledge, attitudes, views and perspectives of CoP examiners regarding standard setting were evaluated 18 months after standard setting was introduced for the written components of the Fellowship examinations.

For ease of reference the relevant research question and the related research objectives, as stated in Chapters 1 and 3, are provided again:

Research Question 2: What are the knowledge, attitudes, views and perspectives of CoP examiners regarding standard setting, and do they change with training and exposure to a process of standard setting?

Research objective 1.4.3.2: Determine the knowledge, attitudes, views and perspectives of the CoP examiners about standard setting.

Research objective 1.4.3.3: Design, deliver and evaluate the impact of a seminar dealing with standard setting in the CoP.

The results of the online survey and the survey conducted before and after delivering the seminar are reported separately, followed by a joint discussion of the findings.

4.2 SURVEY RESULTS

As outlined in section 3.4.1.4 of Chapter 3, the online questionnaire survey of the knowledge, attitudes, views and perspectives of the CoP examiners about standard setting focused on determining:

1. The *initial position* of the CoP examiners in February 2013 (Time 1), 18 months after introducing standard setting in the CoP;
2. The *current position* of CoP examiners *in February 2014* (Time 2), 30 months after introducing standard setting in the CoP; and
3. A *change*, if any, in the knowledge, attitudes, views and perspectives of CoP examiners between Time 1 (T1) and Time 2 (T2) of the survey.

The results of the online survey are reported in the same order as described in Table 3.1 in Chapter 3.

1. Online survey *response rates*
2. Examiners' *knowledge* about standard setting
3. Examiners' prior *education and training* regarding standard setting
4. Examiners' *awareness* of the implementation of standard setting in the CoP
5. *Attitudes, views and perspectives* of examiners regarding standard setting
6. Examiners' opinions about the *feasibility and sustainability* of using the *Angoff method* of standard setting for the CoP examinations.

The reader is advised that the *optional* free text comments box provided in the online survey was seldom used by respondents and yielded scanty data, which were insufficient for further analysis and, therefore, not explored further.

4.2.1 Online survey response rates

As described in Chapter 3, the online survey was sent to *all* the members of the CoP who were appointed as examiners at the commencement of the study (2010 – 2012), during *both* data collection rounds. These examiners were called the 'current' CoP examiners (n=54). Of these 54 examiners, 38 (70.4%) responded to the first round of the survey conducted in February 2013 (T1), and 41 (75.9%) responded to the second round of the survey conducted in February 2014 (T2). There were 33 (61.1%) examiners who responded to both rounds of the survey (Table 4.1). The survey software package calculated the margin of error in the results in the population of 54 CoP examiners as less than 10%. The exact margins for each survey round are reported in Table 4.1.

TABLE 4.1: RESPONSE RATES FOR THE ONLINE SURVEY

Parameter	Time 1 Feb 2013	Time 2 Feb 2014
Participants who completed the survey **twice**	33 (61%)	33 (61%)
Participants who completed the survey **once**	5 (9.3%)	8 (14.8%)
Overall response rate	38 (70.4%)	41 (75.9%)
Margin of error in results*	8.8%	7.7%

*calculated by the formula provided with survey software (SurveyMonkey Inc. 2014)

4.2.2 Knowledge about standard setting

As explained in Chapter 1, the Cohen method of standard setting was introduced and incrementally implemented in the written components of the Fellowship examinations of the CoP starting in August 2011. Therefore, the CoP examiners were asked to report on their self-perceived knowledge about the concept of standard setting in general, and the Cohen method more specifically. Although the examiners, as part of the research project design also participated in experimental standard setting processes using the Angoff method, they were not expected to be knowledgeable about the Angoff method other than a basic understanding of the procedure, as used in the study, and a clear understanding of the meaning of the term 'the borderline student'.

4.2.2.1 *Knowledge about the concept of standard setting in general*

The overall self-reported data of the CoP examiners regarding their general knowledge on the concept of standard setting, from both rounds of the survey, are presented in Figure 4.1 below. The data from the lowest two knowledge levels "I know *nothing* about it" and "I know *very little* about it" were combined from the survey results, since the numbers in each category were very low and they were deemed to measure similar knowledge levels.

Table 4.2 provides the self-reported, matched or paired data for the 33 examiners who completed both rounds on the survey (T1 and T2). The Marginal Homogeneity Omnibus Test was used to detect statistically significant changes across the self-reported knowledge levels of the matched data in this section. It reported that

statistically significant changes between T1 to T2 were present in one or more self-reported knowledge levels (p=0.001). The results for each of the four levels are reported below.

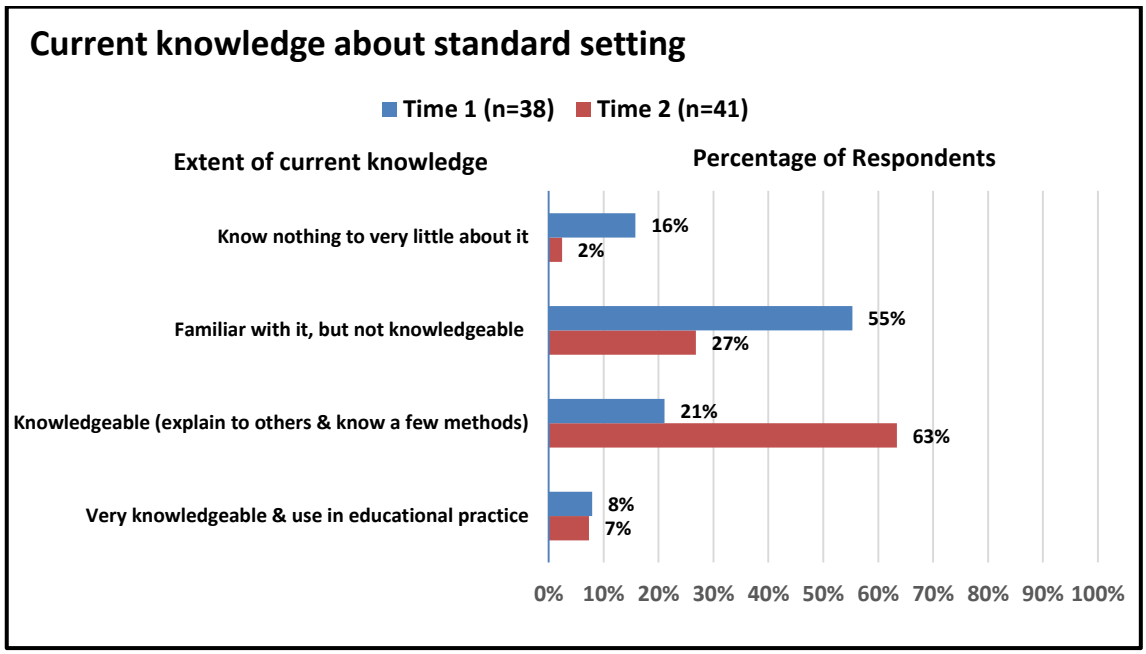


FIGURE 4.1: SELF-REPORTED KNOWLEDGE ON THE CONCEPT OF STANDARD SETTING
(Compiled by the researcher, SCHOEMAN 2014)

Figure 4.1 shows that the general conceptual knowledge of the CoP examiners, about standard setting, increased from T1 to T2. In T1 of the survey, 16% of examiners (n=6) reported they knew 'nothing to very little' about the concept of standard setting; at T2 it was down to 2% (n=1) (*cf.* Figure 4.1). The paired data in Table 4.2 confirms the trend, but the change in this category, from T1 (n=5, 15%) to T2 (n=0), was not statistically significant (p=0.063).

The data in Figure 4.1 shows that the most frequent response at T1, "familiar, but not knowledgeable", dropped from 55% (n=21) to 27% (n=11) at T2. The matched data in Table 4.2 shows a similar trend but the difference was not statistically significant (p= 0.049) due to the Bonferroni adjustment in the alpha level for multiple comparisons (see footnote under Table 4.2 for explanation of Bonferroni adjustment).

Figure 4.1 shows that the predominant self-reported knowledge level at T2 was "knowledgeable and *able to explain* it to a colleague and know a few methods" (n=26, 63%). This level was selected by eight examiners (21%) at T1 and increased to 63%

(n=26) at T2. The matched data in Table 4.2, confirms that this change was statistically significant, even after the Bonferroni adjustment (p=0.001).

Similar knowledge levels were reported at the top-end of the scale in both rounds (*cf.* Figure 4.1 and Table 4.2).

TABLE 4.2: SELF-REPORTED KNOWLEDGE ON THE CONCEPT OF STANDARD SETTING (MATCHED DATA, n=33 - 61.1 % of CoP examiners)

Standard setting – current knowledge	N (%)		*p*-value*
	T1	T2	
I know nothing to very little about it	5 (15%)	0 (0%)	0.063
I am familiar with it, but not knowledgeable about it	18 (55%)	9 (27%)	0.049
I am knowledgeable about it – I can explain it to a colleague and know a few methods	8 (24%)	22 (67%)	*0.001*
I am very knowledgeable about it and use standard setting methods in my own educational practice	2 (6%)	2 (6%)	no change

* Post-hoc McNemar tests used with a Bonferroni adjustment in the alpha level for four comparisons (0.05/4) = 0.0125.

Red italic p-value indicates statistically significant difference in this level from T1 to T2

4.2.2.2 *Knowledge about the Cohen method of standard setting*

The overall data from both rounds of the survey regarding the CoP examiners' self-reported knowledge on the Cohen method of standard setting are presented in Figure 4.2 below. Data from the lowest two knowledge levels "I have *no idea* how it works" and "I have *a vague idea* how it works, but can't explain it to a colleague with confidence" were combined from the survey results, since they were deemed to measure similar levels of knowledge.

Table 4.3, provides the self-reported, matched or paired data for 33 examiners who completed both rounds on the survey (T1 and T2). The Marginal Homogeneity Omnibus Test was used to test holistically for statistically significant changes across the three self-reported knowledge levels of the matched data in this section. Statistically significant changes between T1 and T2 did occur in one or more self-

reported knowledge levels (p=0.001). The results for each of the three levels are reported below.

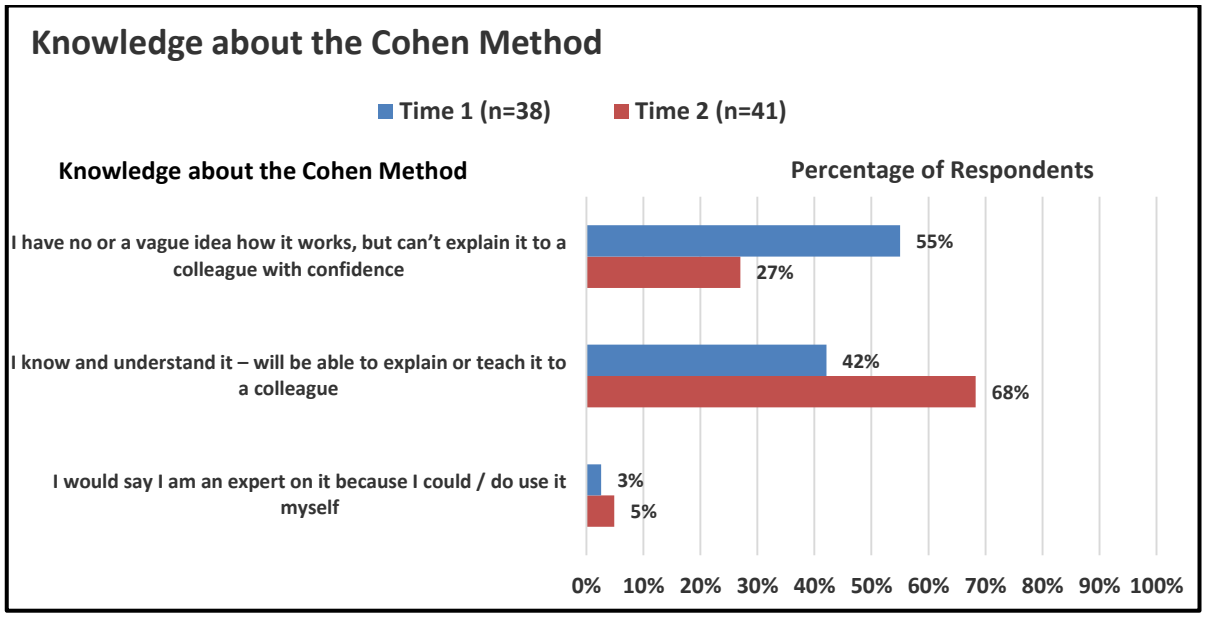


FIGURE 4.2: OVERALL SELF-REPORTED KNOWLEDGE ABOUT THE COHEN METHOD
(Compiled by the researcher, SCHOEMAN 2014)

Overall, the data from Figure 4.2 indicates a positive shift of increased self-reported knowledge about the Cohen method from T1 to T2. The initial combined "I have no or a vague idea about the Cohen method" levels of T1 respondents (n=22, 55%) dropped by 28% at T2 (n=11, 27%). In contrast, the "I know and understand it and will be able to *explain or teach* it to a colleague" level increased by 26% from T1 (n=16, 42%) to T2 (n=28, 68%).

The matched data in Table 4.3 mirrors the changes which took place between T1 and T2 in Figure 4.2. There were statistically significant changes for both levels of self-reported knowledge of the Cohen method, with a clear shift towards the "I know and understand it – will be able to explain or teach it to a colleague" level. The respective p-values for each level are reported in Table 4.3.

The number of self-reported experts on the Cohen method remained virtually unchanged from T1 (n=1) to T2 (n=2) (*cf.* Figure 4.2) and the matched data (Table 4.3) confirmed that the change between T1 and T2 was not statistically significant (p=1.000).

TABLE 4.3: SELF-REPORTED KNOWLEDGE ON THE COHEN METHOD (MATCHED DATA, n=33 - 61.1 % of CoP examiners)

Cohen method – current knowledge	N (%)		*p*-value*
	T1	T2	
I have no a vague idea how it works, but can't explain it to a colleague with confidence	18 (55%)	8 (24%)	*0.002*
I know and understand it – will be able to explain or teach it to a colleague	14 (42%)	23 (70%)	*0.012*
I would say I am an expert on it because I could / do use it myself	1 (3%)	2 (6%)	1.000

* Post-hoc McNemar tests used with a Bonferroni adjustment in the alpha level for three comparisons (0.05/3) = 0.0167.

Red italic p-value indicates statistically significant difference in this level from T1 to T2

4.2.3 Education and Training on standard setting

In this section, examiners could select all options which applied to them. The most important change that took place from T1 to T2 was observed on the lower end of the scale. The number of examiners who had not had any education or training about standard setting decreased from ten (26%) at T1 to two (5%) at T2 (*cf.* Figure 4.3). In addition, the overall data on attendance at one or more workshops/seminars on standard setting increased from T1 (n=22, 61%) to T2 (n=33, 81%). Table 4.4 provides the matched data for this section and although there were no statistically significant changes on any level, the trend mirrored that seen in Figure 4.3.

There were small insignificant changes in the overall cohort and matched data between T1 and T2 at the upper levels of the scale on education and training (*cf.* Figure 4.3 and Table 4.4).

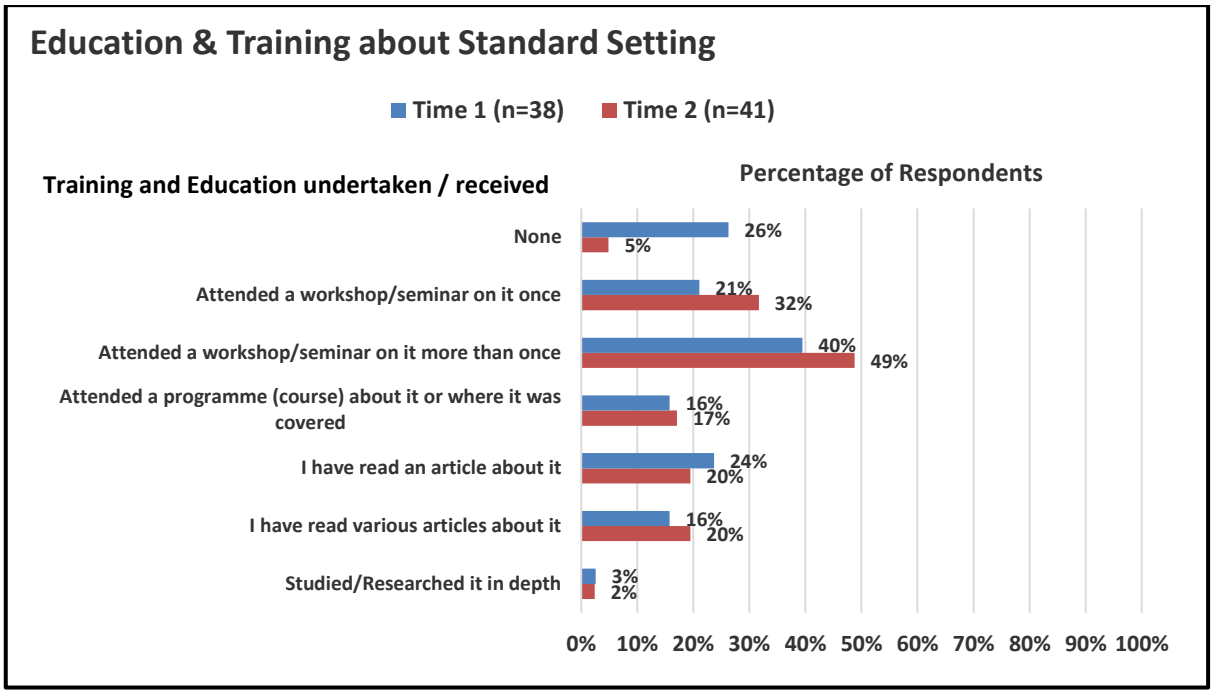


FIGURE 4.3: EDUCATION AND TRAINING ABOUT STANDARD SETTING (Compiled by the researcher, SCHOEMAN 2014)

TABLE 4.4: EDUCATION AND TRAINING ON THE CONCEPT OF STANDARD SETTING (MATCHED DATA, n=33 - 61.1 % of CoP examiners)

Education & Training about Standard Setting	N (%)		*p*-value*
	T1	T2	
None	7 (21%)	1 (3%)	0.031
Attended a workshop/seminar on it once	7 (21%)	11 (33%)	0.344
Attended a workshop/seminar on it more than once	14 (42%)	17 (51%)	0.581
Attended a programme (course) about it or where it was covered	6 (18%)	7 (21%)	1.000
I have read an article about it	9 (27%)	6 (18%)	--
I have read various articles about it	6 (18%)	7 (21%)	1.000
Studied / Researched it in depth	1 (3%)	1 (3%)	no change

* Post-hoc McNemar tests used with a Bonferroni adjustment in the alpha level for seven comparisons (0.05/7) = 0.007.

4.2.4 Awareness of implementation of standard setting

At Time 1, in February 2013, standard setting (the Cohen method) had not yet been introduced for the FCP (SA) Part II SEQ test. As a result, the examiners were only asked, in both rounds, about their awareness of standard setting changes implemented

in the FCP (SA) written test formats at T1 of the survey, which were the Part I MCQ test and the Part II Objective test.

TABLE 4.5: OVERALL AWARENESS OF THE INTRODUCTION OF THE COHEN METHOD IN THE CoP

Awareness of introducing the Cohen method		**N (%)**	
		T1	**T2**
2011 introduction for the **Part I MCQ test**	YES	30 (79%)	39 (95%)
	NO	8 (21%)	2 (5%)
2012 introduction for the **Part II Objective Test**	YES	21 (55%)	39 (95%)
	NO	17 (45%)	2 (5%)

Table 4.5 provides the overall data regarding the awareness of the CoP examiners about the introduction of the Cohen method of standard setting in the CoP. The data at T1 indicated that awareness of the introduction of the Cohen method for the MCQ test was noticeably higher (n=30, 79%) than for the Objective Test (OT) (n=21, 55%). Approximately half of the CoP examiners at T1 (45%, n=21) were unaware that standard setting had been introduced for the OT in 2012. This was in contrast to only eight examiners (21%) who did not know about the introduction of standard setting for the MCQ test in 2011 (*cf.* Table 4.5). Awareness about the implementation of the Cohen method for the MCQ test and the OT improved from T1 to *identical* levels for both formats in T2 (95%, n=39).

TABLE 4.6: AWARENESS OF THE INTRODUCTION OF THE COHEN METHOD IN THE CoP (MATCHED DATA, n=33 - 61.1 % OF CoP EXAMINERS)

Awareness of introducing the Cohen method	**N (%)**		***p*-value***
	T1	**T2**	
2011 introduced for the **Part I MCQ test** (YES data)	28 (85%)	32 (97%)	0.219
2012 introduced for the **Part II OT** (YES data)	18 (55%)	32 (97%)	*<0.001*

* McNemar tests used alpha = 0.05.

Red italic p-value indicates statistically significant difference in this item from T1 to T2

This pattern was similar for the matched data provided in Table 4.6, which indicated that awareness of the use of the Cohen method for the Part I MCQ test was high at T1 and T2 (p=0.219). However, there was a statistically significant (p<0.001) increase in awareness of the use of the Cohen method for the Part II OT from T1 (n=18, 55%) to T2 (n=32, 97%).

4.2.5 Attitudes, views and perspectives of standard setting

In this section, the data from the survey items relating to the attitudes, views and perspectives about standard setting of the CoP examiners are presented and discussed. The data are presented in the three sub-section topics of section 4 of Table 3.1, which were:

1. *Changing* the traditional past practice of using a fixed 50% pass mark for all the written components of the FCP (SA) examination
2. The *current use* of the Cohen method for two written papers of the FCP (SA) examination – the Part I MCQ and Part II Objective Test papers
3. The possible *expanded use* of the Cohen method in the future to include the final written component in the FCP (SA) examination, the Part II SEQ test.

The responses to these questions were subsequently classified using change management terms, depending on whether they were *'Adopters'* (who supported change) or *'Non-adopters'* (who were either *uncertain* or *resisting* change). The data from the ten survey items addressing these three broad 'topics of change' were grouped together, to facilitate triangulation of the data regarding the attitudes, views and perspectives of the CoP examiners at T1 and T2. The coding procedure used to cluster the data from Items 7 – 16 of the survey into the format of Table 4.7 and 4.8, is provided in Appendix D.

The data are presented in two tables:
Table 4.7 contains the *responses* for the all the participants at T1 (n=38) and T2 (n=41). This is regarded as the *full* situational analyses data at time T1 and T2 of the survey.
Table 4.8 contains the *matched* data of the examiners who completed *both* rounds of the survey (n=33). This was done to investigate any changes or potential shifts in 'hearts and minds' of the CoP examiners between T1 and T2.

The layout of Table 4.7 and Table 4.8 were kept similar to facilitate easy reading and understanding of the data presented.

TABLE 4.7: SITUATIONAL ANALYSES - VIEWS, ATTITUDES AND PERSPECTIVES OF CoP EXAMINERS REGARDING STANDARD SETTING: UNMATCHED DATA

			Adopters					Non-adopters									
			Supporting change					Uncertain					Resisting change				
Item no.	Item Type	Topic of change	Time 1 (n=38)		Time 2 (n=41)		T1-T2 shift	Time 1 (n=38)		Time 2 (n=41)		T1-T2 shift	Time 1 (n=38)		Time 2 (n=41)		T1-T2 shift
			n	%	n	%	%	n	%	n	%	%	n	%	n	%	%
Q8	MCQ	Change previous fixed 50% pass mark in Part I MCQ	21	55.3%	32	78.0%	22.8%	6	15.8%	4	9.8%	-6.0%	11	28.9%	5	12.2%	-16.8%
Q9	MCQ	Change previous fixed 50% pass mark in Part II OT	20	52.6%	32	78.0%	25.4%	7	18.4%	4	9.8%	-8.7%	11	28.9%	5	12.2%	-16.8%
Q14	Likert	Change previous fixed 50% pass mark in all FCP exams	30	78.9%	34	82.9%	4.0%	5	13.2%	2	4.9%	-8.3%	3	7.9%	5	12.2%	4.3%
		Mean:		**62.3%**		**79.7%**	**17.4%**		**15.8%**		**8.1%**	-7.7%		**21.9%**		**12.2%**	-9.7%
Q10	MCQ	Use of Cohen method in Part I MCQ	32	84.2%	35	85.4%	1.2%	5	13.2%	3	7.3%	-5.8%	1	2.6%	3	7.3%	4.7%
Q11	MCQ	Use of Cohen method in Part II OT	33	86.8%	35	85.4%	-1.5%	4	10.5%	3	7.3%	-3.2%	1	2.6%	3	7.3%	4.7%
Q12	Likert	Use standard setting in all FCP exams	29	76.3%	35	85.4%	9.1%	5	13.2%	2	4.9%	-8.3%	4	10.5%	4	9.8%	-0.8%
Q13	Likert	Endorse the current use of the Cohen method (MCQ & OT)	29	76.3%	35	85.4%	9.1%	5	13.2%	2	4.9%	-8.3%	4	10.5%	4	9.8%	-0.8%
		Mean:		**80.9%**		**85.4%**	**4.4%**		**12.5%**		**6.1%**	-6.4%		**6.6%**		**8.5%**	2.0%
Q7	MCQ	Expanded use of Cohen method in SAQ	23	60.5%	32	78.0%	17.5%	10	26.3%	2	4.9%	-21.4%	5	13.2%	7	17.1%	3.9%
Q15	Likert	Expanded use of Cohen method in SAQ	26	68.4%	34	82.9%	14.5%	5	13.2%	1	2.4%	-10.7%	7	18.4%	6	14.6%	-3.8%
Q16	Likert	Expanded use of Cohen method in SAQ	28	73.7%	37	90.2%	16.6%	8	21.1%	0	0.0%	-21.1%	2	5.3%	4	9.8%	4.5%
		Mean:		**67.5%**		**83.7%**	**16.2%**		**20.2%**		**2.4%**	-17.7%		**12.3%**		**13.8%**	1.5%
		All items Mean:		**71.3%**		**83.2%**	11.9%		**15.8%**		**5.6%**	-10.2%		**12.9%**		**11.2%**	-1.7%
		All items SD		**11.8%**		**4.1%**			**4.8%**		**3.1%**			**9.7%**		**3.1%**	
		Topic means min/max:		62-81%		**80-85%**			12-20%		**2-8%**			6-22%		**8-14%**	
		Year:		2013		2014			2013		2014			2013		2014	

TABLE 4.8: SITUATIONAL ANALYSES - VIEWS, ATTITUDES AND PERSPECTIVES OF CoP EXAMINERS REGARDING STANDARD SETTING: MATCHED DATA (n=33, 61.1 % of CoP examiners)

				Adopters						Non-adopters											
				Supporting change						Uncertain						Resisting change					
Item no.	Item Type	Topic of change	Overall *p*-value	Time 1		Time 2		T1-T2 shift	*p*-value	Time 1		Time 2		T1-T2 shift	*p*-value	Time 1		Time 2		T1-T2 shift	*p*-value
				n	%	n	%	%		n	%	n	%	%		n	%	n	%	%	
Q8	MCQ	Change previous fixed 50% pass mark in Part I MCQ	0.008*	18	54.5%	26	78.8%	24.2%	0.021*	6	18.2%	4	12.1%	-6.1%	-	9	27.3%	3	9.1%	-18.2%	0.070
Q9	MCQ	Change previous fixed 50% pass mark in Part II OT	0.005*	17	51.5%	26	78.8%	27.3%	0.012*	7	21.2%	4	12.1%	-9.1%	-	9	27.3%	3	9.1%	-18.2%	0.070
Q14	Likert	Change previous fixed 50% pass mark in all FCP exams	0.593	27	81.8%	28	84.8%	3.0%	-	5	15.2%	1	3.0%	-12.1%	-	1	3.0%	4	12.1%	9.1%	-
		Mean:			**62.6%**		**80.8%**	**18.2%**			**18.2%**		**9.1%**	**-9.1%**			**19.2%**		**10.1%**	**-9.1%**	
Q10	MCQ	Use of Cohen method in Part I MCQ	0.285	30	90.9%	29	87.9%	-3.0%	-	3	9.1%	1	3.0%	-6.1%	-	0	0.0%	3	9.1%	9.1%	-
Q11	MCQ	Use of Cohen method in Part II OT	0.285	30	90.9%	29	87.9%	-3.0%	-	3	9.1%	1	3.0%	-6.1%	-	0	0.0%	3	9.1%	9.1%	-
Q12	Likert	Use standard setting in all FCP exams	0.841	26	78.8%	28	84.8%	6.1%	-	4	12.1%	1	3.0%	-9.1%	-	3	9.1%	4	12.1%	3.0%	-
Q13	Likert	Endorse the current use of the Cohen method (MCQ & OT)	0.827	27	81.8%	28	84.8%	3.0%	-	4	12.1%	1	3.0%	-9.1%	-	2	6.1%	4	12.1%	6.1%	-
		Mean:			**85.6%**		**86.4%**	**0.8%**			**10.6%**		**3.0%**	**-7.6%**			**3.8%**		**10.6%**	**6.8%**	
Q7	MCQ	Expanded use of Cohen method in SAQ	0.758	22	66.7%	26	78.8%	12.1%	-	7	21.2%	1	3.0%	-18.2%	-	4	12.1%	6	18.2%	6.1%	-
Q15	Likert	Expanded use of Cohen method in SAQ	0.384	24	72.7%	28	84.8%	12.1%	-	4	12.1%	1	3.0%	-9.1%	-	5	15.2%	4	12.1%	-3.0%	-
Q16	Likert	Expanded use of Cohen method in SAQ	0.670	25	75.8%	29	87.9%	12.1%	-	6	18.2%	0	0.0%	-18.2%	-	2	6.1%	4	12.1%	6.1%	-
		Mean:			**71.7%**		**83.8%**	**12.1%**			**17.2%**		**2.0%**	**-15.2%**			**11.1%**		**14.1%**	**3.0%**	
		All items Mean:			**74.5%**		**83.9%**	9.4%			**14.8%**		**4.5%**	-10.3%			**10.6%**		**11.5%**	0.9%	
		All items SD			**13.6%**		**3.8%**				**4.6%**		**4.1%**				**10.0%**		**2.8%**		
		: means min/max:			63-86%		**81-86%**				11-18%		**2-9%**				4-19%		**10-14%**		
		Year:			2013		2014				2013		2014				2013		2014		

*** = Statistical significant value**

Unfortunately, it was not possible to compute inferential statistics on the differences between the T1 and T2 'means' data – the three 'Topic of change' means and the 'All items means' in Table 4.8. The reason why comparing 'total' averages, without individual level data is not possible, is because the concept of inferential statistics refers to inferring characteristics of populations from characteristics of samples. T-tests and other similar 'comparing the means' tests, are statistical tests that are carried out on samples so that inferences can be made about populations. Thus, in order to run statistical tests to compare averages, a sample is needed from the population, and not merely the two mean values themselves (Howell 2007:5).

4.2.5.1 *General comments on the attitudes, views and perspectives about standard setting of the CoP examiners*

As illustrated in Table 4.7, across all ten items addressing the three topics of change, the overall mean level of initial support by the 'Adopters' at T1 was high at 71.3%, with a standard deviation (SD) of 11.8%. Uncertainty about the changes made and proposed stood at 15.8% (SD=4.8%) and the level of resistance to change by the 'Non-adopters' was 12.9% (SD=9.7%). This was the summarised initial situational analysis in February 2013, 18 months after the initiation of the change process, i.e. the introduction and implementation of a standard setting method in the CoP. There were 38 participants in this round.

After a further 12 months of exposure to standard setting, and attendance at a customised educational seminar about standard setting in the CoP, based on the learning needs identified at T1, the second situational analysis was conducted in February 2014 (T2), using the same survey instrument and 41 examiners responded.

The position regarding the attitudes, views and perspectives about standard setting in the CoP in February 2014 (T2) is also provided in Table 4.7. The support for change, regarding standard setting, increased on average by nearly 12% to 83.2%, with a reduced SD of 4.1%. The uncertainty level decreased by 10.2% to 5.6% (SD reduced to 3.1%) in at T2 and resistance to change remained largely the same with a small reduction of 1.7% in the 'Non-adopters' group to 11.2% (SD reduced to 3.1%).

Therefore, the general trend across the ten items presented in Table 4.7, from T1 to T2, shows an increase in support, a reduction in uncertainty and similar levels of resistance towards change (the introduction of standard setting). The standard deviations in Table 4.7 all reduced from T1 to T2, which suggests an increase in the levels of convergence and agreement in the opinions of the examiners in T2. Both the general and standard deviation tendencies were replicated by the matched data in Table 4.8.

The results for each of the three topics of change are described separately in sections 4.2.5.2 – 4.2.5.4 below, which include regular referrals to Tables 4.7 and 4.8.

4.2.5.2 *Changing from the traditional fixed 50% pass mark*

Items 8, 9 and 14 of the survey specifically addressed this topic. The overall (unmatched) data for this topic is presented in Table 4.7. The mean support for changing the fixed 50% pass mark, traditionally used in all FCP (SA) written examinations, was 62.3% in T1. This increased by 17.4% in T2 to a mean support level of 79.7% of CoP examiners. The uncertain group showed reductions of about 6-8% across all three items (mean = 7.7%) and ended at T2 with, on average, 8.1% of examiners still uncertain whether changing the previous fixed 50% pass mark was a positive or negative change. The examiners resisting change on this topic also decreased from T1 to T2 by 9.7% to 12.2% on average, although in item 14 a slight increase in resistance of 4% (n=2) was noted at T2.

The matched data in Table 4.8 shows a statistically significant difference (change), between T1 and T2, for Item 8 (p=0.008) and Item 9 (p=0.005). Post-hoc analysis showed the significant change occurred in the 'Adopters' group for both items. Support for changing the previous fixed 50% pass mark for the Part I MCQ test (item 8) and Part II OT (item 9) assessments increased significantly between T1 and T2 by 24.2% (p=0.021) and 27.3% (p=0.012) respectively. The support for changing the previous pass mark in the MCQ test and OT stood equally at 78.8% at T2. Examiners who changed their views, as expressed in Item 8 (n=8) and Item 9 (n=9), came predominantly from the 'Non-adopters (Resisting change)' group (n=6), who were resistant to change during the T1 survey in February 2013 and the remainder came from the 'Uncertain group' (n=2).

4.2.5.3 *Current use of the Cohen method*

Four survey items addressed this topic in the survey. They were Items 10-13. The initial survey conducted in February 2013 (T1) showed a high level of support for the introduction and implementation of the Cohen method in the CoP. At that stage, on average, 80.9% of examiners had adopted the change to using the Cohen method (*cf.* Table 4.7). The matched data in Table 4.8 shows that the level of support for the current use of the Cohen method, in the Part I MCQ test and the Part II OT was high and changed little over time. Both the overall unmatched data (Table 4.7) and the matched data (Table 4.8) showed an increase in support for change of 4.4% (n=6)

and 0.8% (n=1) at T2 respectively. The T2 support levels in both tables were approximately 86%.

The 'Uncertain group' reduced in all four items, in both tables, by about 6-7% on average from T1 to T2, ending with mean uncertainty levels of 6.1% overall (Table 4.7) and 3% in the matched data (Table 4.8) respectively. The 'Resisting change group' showed a mean increase of 2% overall and 6.8% in the matched data to end on 8.5% and 10.6% respectively after T2.

4.2.5.4 *Expanded use of the Cohen method*

At the time of administering the survey in February 2013 (T1), the Cohen method of standard setting had not yet been implemented in the Part II SEQ test. It was the last remaining written component of the FCP (SA) examination subject to a fixed 50% pass mark. Although the CoP implemented the Cohen method for the SEQ papers in May 2013, the decision was made not to alter the survey being used in February 2014 (T2). This was done to allow analysis of any potential changes in the attitudes, views and perspectives of the CoP examiners regarding the suitability of the Cohen method for the SEQ assessment, at T1 initially. Items 7, 15 and 16 from the questionnaire survey explored the CoP examiners' attitudes, views and perspectives on this topic.

As seen in Table 4.8, none of the matched data for the three items, which addressed this topic, showed any statistically significant changes between February 2013 (T1) and February 2014 (T2). This was most likely due to the low numbers of paired examiners who changed their opinions from T1 to T2. However, reviewing this topic's results from the overall data (Table 4.7) provided some useful findings, which are described below.

The initial T1 mean level of support for the expanded use of the Cohen method (in the SEQ test) stood overall at 67.5% (Table 4.7) and the matched data on 71.7% (Table 4.8). After T2, the mean support level in the CoP for expanding the use of the Cohen method to the SEQ test rose to nearly 84% in both tables. This was close to the same level of adoption and support for the Cohen method in its current use.

The uncertain groups in both tables maintained the same downward trend as in the other two topics from T1 to T2. They dropped to approximately 2% uncertainty at T2.

The 'Resisting change group' remained relatively constant at about 12-14% from T1 to T2.

4.2.6 Feasibility and sustainability of the Angoff method in the CoP

At T2 of the survey in February 2014, there was a need to evaluate the feasibility and sustainability of the Angoff process, as used in the CoP as part of this study. If the Angoff method performed well and had a favourable utility in this context, the CoP might wish to consider its formal implementation as a replacement for the Cohen method, going forward. Therefore, it was important to determine if the Angoff method was a viable alternative, from a practicability perspective, to the Cohen method. The CoP examiners, who responded to T2 of the survey and participated in one or more of the five Angoff meetings, were asked to give their opinion about its long term feasibility and sustainability in the CoP.

At T2, 56% (n=23) of the respondents indicated that they had participated at some stage in the Angoff standard setting meetings and processes. The views on this aspect of the Angoff method as used in the CoP are reported in Table 4.9.

TABLE 4.9: FEASIBILITY AND SUSTAINABILITY OF THE ANGOFF METHOD IN THE CoP

The Angoff Method in the CoP: Feasible and Sustainable in the long run?	**T2 responses (n=23) N (%)**
Yes, I think it is feasible and we can sustain it with our resources	2 (9%)
Uncertain	3 (13%)
No, I think it is not feasible and too resource intensive to sustain in the long run	18 (78%)

As illustrated in Table 4.9, 78% of CoP examiners (n=18), who took part in the standard setting process using the Angoff method thought that it was not feasible and sustainable in the long term.

4.3 EDUCATIONAL SEMINAR RESULTS

In total, 24 (44.4%) of the 54 current CoP examiners (*cf.* section 3.4.1.2) attended one of the two customised educational seminars about standard setting held in May and October 2013. Since both seminars were *identical* in format and content, examiners were encouraged to attend one of them. The content of the seminars was based on the learning needs derived from the survey conducted at T1. The slides used in the presentation are attached in Appendix B-1.

The main focus of the seminars was to clarify the rationale, concept and principles of standard setting, explain the methodology of the Angoff and Cohen methods, create awareness of the implementation of the Cohen method in the CoP and present the findings of the first round of the survey completed in February 2013 (T1).

Evaluation of the impact of the standard setting seminar focussed on the pre- and post-seminar understanding and opinions of the CoP examiners regarding the overall *utility* of the Cohen method of standard setting as compared to the Angoff method.

As described in Chapter 3, the examiners were asked to use a 7-point rating scale (1=lowest, 7=highest) to rate each of the seven utility parameters (UPs), for both standard setting methods, based on their current understanding and opinion. The UPs included for rating each method were derived from the literature review and included: objectivity, credibility, validity, reliability, transparency, feasibility and sustainability.

Attendees were also provided with an "I don't know" option on the evaluation sheets. The evaluation sheet used during both seminar evaluations is attached as Appendix B-2. The results of the pre- and post-seminar evaluations are presented below.

4.3.1 Pre-seminar evaluation of understanding and opinion

As can be seen in Table 4.10, the examiners struggled to assign ratings for the seven UPs before attending the seminar. On average, only 23.2% and 51.2% examiners provided a numerical rating in the pre-seminar round for the Angoff and Cohen methods respectively (*cf.* Table 4.10). Their understanding of the Angoff method was particularly limited; 63.7% of examiners used the "I don't know" option for the seven UPs for the Angoff method and 36.3% for the Cohen method (*cf.* Table 4.10). No

ratings (blank spaces) were assigned by 13.1% and 12.5% of examiners for the Angoff and Cohen methods respectively.

TABLE 4.10: EDUCATIONAL STANDARD SETTING SEMINAR FOR THE CoP: PRE-SEMINAR EVALUATION RESULTS (n=24 attendees)

Participant response	**PRE**-seminar results	
	Angoff	**Cohen**
"I don't know" (%)	63.7%	36.3%
No rating given (%)	13.1%	12.5%
Rating given (%)	23.2%	51.2%

Due to the low number of ratings and level of understanding expressed, comparisons between the two methods were not made using the pre-seminar data. The post-seminar evaluation data were seen as the critical information for comparing the opinions and levels of understanding of the two methods by the CoP examiner attendees.

4.3.2 Post-seminar evaluation of understanding and opinion

After the seminar there were no "I don't know" ratings; every examiner had an opinion about the seven UPs of each method, as used in the CoP.

The data produced by the ordinal rating scales used in the evaluation sheets to capture the opinions and understanding of the examiners, were non-parametric in nature and hence the *median* ratings, for each UP, were compared between the two methods. The post-seminar median ratings for each method, per UP, are provided in Table 4.11. There were statistically significant differences noted for *all seven* UPs between the Angoff and Cohen methods using the Wilcoxon signed rank test ($p<0.001$). The Cohen method was consistently rated higher across all seven UPs.

TABLE 4.11: EDUCATIONAL STANDARD SETTING SEMINAR FOR THE CoP: POST-SEMINAR EVALUATION RESULTS (n=24 attendees)

Utility parameter evaluated	**POST**-seminar results Median Ratings (0-7)		*p*-value
	Angoff	**Cohen**	
Objectivity	3	6	*<0.001*
Credibility	4	6	*<0.001*
Validity	4	6	*<0.001*
Reliability	3	6	*<0.001*
Transparency	3	7	*<0.001*
Feasibility	2	6	*<0.001*
Sustainability	2	7	*<0.001*

Red italic p-values indicates statistically significant difference in UP median rating

4.4 DISCUSSION

This chapter reports on the first component of this study, which addressed the question: "What are the knowledge, attitudes, views and perspectives of examiners of the CoP regarding standard setting, and do they change with training and exposure to a process of standard setting?"

The knowledge, attitudes, views and perspectives of the CoP examiners about standard setting were evaluated using an online questionnaire survey in February 2013 (T1) and a year later in February 2014 (T2).

The results of this chapter are discussed in the same order in which they were presented in the preceding text.

4.4.1 Online survey response rates

The good response rates of more than 70% for both rounds of the online survey in February 2013 (Time 1, n=38, 70.4%) and February 2014 (Time 2, n=41, 75.9%) resulted in error margins of 8.8% and 7.7% respectively, for the results of the survey (SurveyMonkey Inc. 2014) (*cf.* Table 4.1). Given the sensitive nature of change management in high-stakes licensing assessment, as explained in Chapter 2, the survey was administered anonymously and *no* demographic information about the CoP

examiners was requested. The results of the surveys were, therefore regarded as representative of the CoP examiners from a *numbers and characteristics* perspective but not necessarily from a demographic composition perspective. The appointment system and basic characteristics of CoP examiners were explained in Chapter 3 (*cf.* section 3.4.1.2).

The potential influence of sampling error from the non-responders is recognised, but its effect is probably minimal, since less than 15% (n=8) of CoP examiners did not respond to any of the survey rounds. There were 33 examiners (61.1% of population) who completed *both* rounds of the survey and the matched data, in particular, were used to identify changes in the CoP examiners between Time 1 (T1) and Time 2 (T2).

4.4.2 Knowledge about standard setting

General conceptual knowledge about standard setting

The CoP examiners had a limited general knowledge about the concept of standard setting at T1. The results from the T2 survey showed that learning had taken place, and the majority of examiners' indicated that they regarded themselves as knowledgeable and able to explain the concept to others. One of the possible contributing factors to this improvement was the fact that 24 examiners (44.4%) attended the CoP educational seminar on standard setting, which was part of this study and specifically aimed at improving the knowledge base regarding standard setting in the CoP. The educational seminar's results are discussed later in this chapter.

Knowledge about the Cohen method

At T1, 55% of the examiners reported that they had no or only a vague idea about how the Cohen method of standard setting works. However, at T2, the situation had changed and more than two thirds of the examiners (68%) indicated that they now knew and understood the Cohen method and that they were able to explain or teach the method to a colleague. This was encouraging from a diffusion of innovation perspective. According to the diffusion model of Rogers, as explained in Chapter 2 (*cf.* section 2.3.7.5) , this meant that the Cohen method was now understood by the "late majority" of CoP examiners. The methodology of the Cohen method was specifically addressed during the educational seminar and hence, might have contributed to these results.

4.4.3 Education and Training about standard setting

While a reduction in the number of examiners who had *no* education or training in standard setting at T1, and an increase in the number who had attended one or more workshops/seminars on standard setting at T2 was not statistically significant, the trend was consistent with the expected outcome of the training and exposure, which took place between T1 and T2. There was no significant improvement between T1 to T2 at the advanced levels of knowledge, which would require more advanced training in the use of standard setting procedures and self-directed initiatives such as reading papers about standard setting.

4.4.4 Awareness of the implementation of standard setting

The time since introduction of the Cohen method may explain the observed difference in the proportion of examiners (24%) who were aware of the implementation of the Cohen method in the Part I MCQ test in 2011 (79%) as compared to the Part II OT in 2012 (55%). However, this situation had changed notably by the T2 survey a year later in February 2014. By that stage, in both test formats, 95% of the examiners were aware of the introduction of standard setting and the matched data showed the change in awareness regarding the introduction of the Cohen method in the OT was statistically significant ($p<0.001$). This was very encouraging from a diffusion of innovation perspective, since nearly all the examiners were now aware of the introduction of the standard setting innovation (the Cohen method) in the CoP. This issue was specifically addressed by the seminar and a difference was anticipated. However, the additional 12 months of exposure to standard setting probably also contributed to the improved awareness of the CoP examiners by T2.

4.4.5 Attitudes, views and perspectives regarding standard setting

General discussion

The similarities between the cohorts' overall situational data analysis (Table 4.5) and the matched data analysis (Table 4.6) is evident and expected, given the relative large number of examiners who completed both rounds (n=33, 61.1%). Support for change in all three 'topic of change' areas increased from T1 to T2 by approximately 12% on average, from an already high mean baseline level of support of 71% to 83%. The high levels of initial adoption and support were unexpected, but encouraging and

indicated that examiners were broadly in favour of the changes/improvements in the assessments. It gave the CoP Council some early reassurance and confidence that the change of assessment practices was widely accepted even before further education and training had taken place.

The data from T2 in February 2014 showed that most of the new 'Adopters' of the changes came from the 'Uncertain group' of examiners, which decreased by 10% on average from T1 to T2. This suggests that the standard setting seminar and further exposure to standard setting procedures addressed the uncertainties of some examiners and convinced them to become 'Adopters' of the standard setting changes in the CoP.

Another general finding was the noted reductions in the standard deviations of the 'All Items Mean' from T1 to T2 in both Tables 4.7 and 4.8. The narrower response ranges in T2 indicated a greater level of agreement and convergence between CoP examiners in T2. It also supports the probable increased levels of education about standard setting in the CoP.

Uncertainty levels dropped from T1 to T2 across all areas/topic/items. This was a further indication that education had taken place and that examiners were less unsure about the standard setting changes in the CoP. The number of examiners still falling into this category at T2 was a mere 5.6% on average.

Overall from T1 to T2, the 'Resisting change' group remained relatively consistent at about 11-13% on average, across all the items. Interestingly, the matched data in Table 4.8 shows that resistance increased in the majority of items, but the absolute number of examiners who were still resisting change remained small ($n \leq 6$). There was, however, a noted *reduction* in T2 of about 17-18% in resistance to changing the previous fixed 50% pass mark, which was a positive development.

Changing the traditional fixed 50% pass mark

There was a high mean level of adoption and support already noted during the T1 survey for changing the previous fixed 50% pass mark practice. Overall, about 62% examiners supported the move in T1, which increased to nearly 80% in T2. This finding was the only statistically significant change recorded during the two survey rounds for the ten items evaluating the 'hearts and minds' of the CoP examiners. This

was particularly encouraging because it was probably the most important change introduced in the CoP's assessment practices to date.

Current use of the Cohen method

The high level of support expressed by the CoP examiners, regarding the use of the Cohen method of standard setting in the MCQ test and OT, was impressive. About 81% of examiners endorsed the use of the Cohen method for the MCQ test and OT during the T1 survey, which increased further to approximately 85% in T2. While the increase was not statistically significant, the observation was very encouraging and reflected widespread agreement with the CoP council's decision to introduce the Cohen method for the MCQ test and OT. A small number of examiners remained uncertain or were resistant to the current use of the Cohen method at both T1 and T2.

The expanded use of the Cohen method

The data show that the CoP examiners strongly supported the expanded use of the Cohen method to include the Part II SEQ test. T1 support levels were about 68% and this increased to approximately 84% at T2. This increase in support may have been influenced by the actual implementation of the Cohen method for the Part II SEQ test in May 2013. In retrospect the decision of the Council was certainly supported and endorsed from the examiners' perspective according to the findings from this study. The number of 'Non-adopters' who remained uncertain about, or resistant to, the expanded use of the Cohen method was very small at both T1 and T2.

4.4.6 Feasibility and sustainability of the Angoff method

More than three quarters of the CoP examiners who had experience of using the Angoff standard setting method, as part of this study, did not think it was a feasible and sustainable long term standard setting alternative (78%). This result was recorded despite conducting the Angoff method, in this study, in the most resource-efficient manner possible (Yes/No version used, with no iterative discussions).

4.4.7 Seminar about standard setting in the CoP - Evaluation

Pre-seminar evaluation

The findings of the pre-seminar evaluation showed there was a real need for more information about standard setting in the CoP. The attending examiners did not have

the knowledge or understanding to form an opinion regarding the utility of the Angoff method, as evidenced by the majority "I don't know" response (64%), and a further 13% who gave no response during the pre-seminar evaluation. The responses were marginally better for the Cohen method since, on average, 51% provided a rating on the UPs and 49% either did not know or gave no rating. These findings triangulated and aligned with the T1 survey results regarding the low levels of knowledge and education on the general concept of standard setting and the Cohen method.

Post-seminar evaluation

The seminar's aims were described previously during its results section. However, it is important to note that the seminar participants were asked to complete the post-seminar evaluation, which was identical to the one used before the seminar, *before* the researcher shared his views about the UPs of the Angoff and Cohen methods.

During the post-seminar evaluation, *all* attending examiners provided an opinion about the UPs of the two methods. This indicated that some education had taken place, even if it only enabled them to move away from the highly prevalent "I don't know" or no answer position before the seminar.

After the seminar the Cohen method was rated significantly higher on all UPs by the participants. The median ratings of the Cohen method were very high in all UPs (6 or 7 throughout), which signified confidence, support and endorsement of the method over the seven UPs.

This was in contrast to the Angoff method, where median ratings were less than half on the 7-point scale (2-3), across five of the seven UPs and only reached the halfway level (4) in 'Credibility' and 'Validity'. This was in line with the literature's perspective that the strongest UPs of the Angoff method were its widely accepted credibility, as a standard setting method, which yielded valid outcomes (Barman 2008:959; Kane 1994:440). It is probably also the reason why it is the most widely used and researched method in general and medical education (Brandon 2004:80). However, the particular Angoff process employed in this study, mostly determined by financial, human and time resources, negativity influenced the examiners' views of the overall utility of the method (Berk 1986:143; Norcini & Guille 2002:818).

The specific Angoff method used in the CoP could be deemed sub-optimal in some aspects since there were no iterations in rating and discussion rounds, and no 'reality check' data. There were, however, more than enough expert judges on the panels and as such, it was the best the CoP could deliver. Even with this 'basic' or purist Angoff strategy, the method was not deemed feasible and sustainable by most of the CoP examiners. Interestingly, the practicability (feasibility and sustainability) were the UPs where the greatest differences were noted between the Angoff and Cohen methods in the post-seminar evaluation. This was not surprising given the cost-effectiveness, time-efficiency and ease of calculation of the Cohen method compared to the resources (time, money and examiners) required by the Angoff method.

4.5 CONCLUSION

This chapter presented the results from the first component of this study which investigated the knowledge, attitudes, views and perspectives of the CoP examiners regarding the introduction and implementation of standard setting in the CoP. Subsequently, a discussion was offered relating to the presented findings. In addition, this chapter also contributes to the overall discussion presented in Chapter 8, where conclusions drawn from this component of the study are provided. In the next chapter, Chapter 5, the results of the second component of the study, specifically relating to the standard setting of the Part I MCQ test, is presented and discussed.

CHAPTER 5

FCP (SA) PART I MULTIPLE CHOICE QUESTION (MCQ) TEST

COMPARING THE PERFORMANCE OF THE ANGOFF AND COHEN METHODS

5.1 INTRODUCTION

In this chapter, the results of the second component of the study, relating specifically to the Part I MCQ test data, are reported and discussed.

For this part of the study, the results of five consecutive cycles of the written, entry-level FCP (SA) Part I MCQ test were analysed and compared. The results reported here contribute towards answering and addressing the research question (see below) and its related objectives, as discussed in Chapter 3.

Research Question 3: Is using the Cohen method, as compared to Angoff method, a more appropriate way of determining the pass mark for the written assessments of the FCP (SA) examination (Part I MCQ test, Part II Objective test and Part II SEQ test)?

Related objective 1.4.3.4: Determine the performance of the Angoff method of standard setting using five cycles of written FCP (SA) examinations data.

Related objective 1.4.3.5: Determine the performance of the Cohen method of standard setting using the same five cycles of written FCP (SA) examinations data as in 1.4.3.4.

Related objective 1.4.3.6: Determine the variability of the scores of the top-performing candidates, in order to substantiate the assumption within the Cohen method that the test scores of the top candidates sitting the FCP (SA) written examinations, are stable and, therefore, comparable.

Related objective 1.4.3.7: Use the findings of Objectives 1.4.3.2 – 1.4.3.6 to contribute to the evaluation of the utility (as defined in Chapter 2) of the Cohen method, as compared to the Angoff method, for the written FCP (SA) examinations.

Summary of methods

A comprehensive description of the methods used in this study was provided in Chapter 3. For convenience, a brief summary is again provided here, mostly to outline the research approach and the data included in the study.

A comparative study design (Altman 1991:6) was used to evaluate the Angoff and Cohen methods of standard setting to answer Research Question 3 and the four related objectives. Five cycles of MCQ test results, a component of the written FCP (SA) examinations, were analysed. Table 3.2 in Chapter 3 describes all the cycles of test data which were included in the study. An abbreviated version of Table 3.2, specific to the test and cycles evaluated in this chapter, is provided to enable easy reference (Table 5.1).

TABLE 5.1: FCP (SA) PART I MCQ TEST DATA INCLUDED IN THE STUDY

Examination name	Format/Type	Exam cycles
Part I MCQ test (entry exam)	**150 MCQ items** selected-response items, single best answer from 5 possible options (with 30 tracker items repeated in all 5 cycles)	Mar 2012
		Aug 2012
		Jan 2013
		Jun 2013
		Feb 2014

The data from the five MCQ test cycles, mentioned in Table 5.1 above, were analysed and are presented in the following order:

- The **candidates' performance data** for all the items in each MCQ test as well specifically for the tracker items to evaluate the quality and comparability of the papers
- The **Angoff standard setting** outcomes for the MCQ test papers
- The **Cohen standard setting** outcomes for the MCQ test papers

5.2 THE PART I MCQ TEST RESULTS

As explained in Chapters 1 and 3, the five cycles of consecutive MCQ tests included in this study all contained 150 best-of-five MCQ items (*cf.* Table 5.1). Thirty of the MCQs in each test were identical and they were used as 'tracker' items so that the performance of consecutive cohorts of candidates could be compared. Due to a technical error only 26 of these items were included in the January 2013 paper. For this test, the score out of 26 was converted to a score out of 30, to enable comparison with the data of the other four cycles.

Cohort sizes

A total of 582 candidates wrote the MCQ tests included in this study. The range of candidates across the five cycles was 80 (71 - 151), with a mean number of candidates per cycle of 116 and a median of 137.

Distribution of data

The performance data across all five MCQ cycles (all 150 items, including the tracker items) were normally distributed according to the Shapiro-Wilk test and Normal Q-Q plot analyses.

The performance data for all the test items (n=150) is presented first, followed by the data for the 30 tracker items.

5.2.1 MCQ test performance data - all 150 items

Descriptive statistics (from Table 5.2)

The *maximum* scores for the five MCQ test cycles were in a narrow range between 78 - 82%, with a median of 79%. In contrast, the *minimum* scores were in a wider range of between -2% and 18%, with a median of 14%.

The *median* of the mean test scores of all five MCQ papers was 46.2% (range= 43.6 – 50.0%). An ANOVA analysis of the mean scores of the five MCQ cycles revealed a statistically significant difference between the mean scores (p=0.008). Post-hoc Tukey's tests showed that there were statistically significant differences between the mean performance of candidates in the *January 2013* paper (43.6%) as compared to the August 2012 paper (49.4%, p=0.038) and the June 2013 paper (50.0%, p=0.018).

The *standard deviation* of the mean scores, across the five cycles, ranged between 12.9 - 15.9% (3%), with a median of 14.6%. The *95% confidence intervals* (CIs) for the mean scores of the cycles are provided in Table 5.2. The width of the CIs for the five MCQ cycles ranged from 4.9 – 6.3%, with a median of 5.2%.

5.2.1.1 *MCQ item analysis - all 150 items*

Proportion Correct value (PC-value)

The overall mean PC-value for all 150 items combined was 0.55. The mean PC-values for each MCQ test ranged between 0.52 - 0.57, with a median of 0.55 (*cf.* Table 5.2). An ANOVA analysis across the five mean PC-values showed *no* statistically significant difference (p=0.959).

Discrimination Index (DI)

The mean DI for each of the five MCQ tests ranged from 0.23 to 0.30, with a median of 0.27 (*cf.* Table 5.2). The mean DI for *all* 750 MCQ items combined (from the five cycles included in the study) was also 0.27.

Item Quality Index (IQI)

The MCQ Item Quality plots are shown in Figure 5.1(a-f). The test items inside the green zone were defined as high-quality test items; they had DI's of 0.20 or more and PC-values, expressed as a percentage, between 20% to 80%. Please refer to Appendix E for a detailed explanation of the interpretation of an Item Quality plot. The IQI for each MCQ test, as derived from the respective Item Quality plots in Figure 5.1(a-e), is provided in Table 5.2. The percentage of high quality MCQs, as previously defined, in each of the five MCQ tests included in the study, ranged from 55 - 69%, with a median value of 59%. Overall, 61% of the 750 test items studied, as seen in Figure 5.1(f), were of a good quality.

TABLE 5.2: THE PART I MCQ TEST - PERFORMANCE DATA (ALL ITEMS)

Part I MCQ exam - All items (150)		Cycles (n= candidates)					Analysis of 5 cycles			
		Mar2012	Aug 2012	Jan2013	Jun2013	Feb2014				
Analysis Component	**Analysis Descriptor**	n= 137	n= 151	n= 80	n= 143	n= 71	**Min**	**Max**	**Range**	**Median**
Descriptive statistics	Maximum score (%)	79	79	78	82	80	78	82	4	79
	Minimum score (%)	-2	14	8	18	18	-2	18	20	14
	Mean score (%)	46.2	49.4	43.6	50.0	45.7	43.6	50.0	6.4	46.2
	Standard Deviation (SD) (%)	14.6	15.4	14.2	15.9	12.9	12.9	15.9	3.1	14.6
	95% confidence intervals (CI) of mean (%)	43.8 - 48.7	46.9 - 51.9	40.4 - 46.7	47.4 - 52.6	42.7 - 48.7	-	-	-	-
	95% CI width (%)	4.9	5.0	6.3	5.2	6.0	4.9	6.3	1.4	5.2
Item analysis	Mean Item Difficulty (PC -value)	0.55	0.57	0.52	0.57	0.54	0.52	0.57	0.06	0.55
	Mean Item Discrimination Index (DI)	0.27	0.29	0.26	0.30	0.23	0.23	0.30	0.07	0.27
	**Item Quality Index (IQI) (%)	58	62	59	69	55	55.0	69.0	14.0	59.0
Test reliability analysis	Cronbach's alpha coefficient	0.92	0.93	0.91	0.93	0.89	0.89	0.93	0.04	0.92
	Standard Error of Measurement (%)	4.2	4.2	4.3	4.2	4.2	4.2	4.3	0.1	4.2

** See text for explanation of IQI

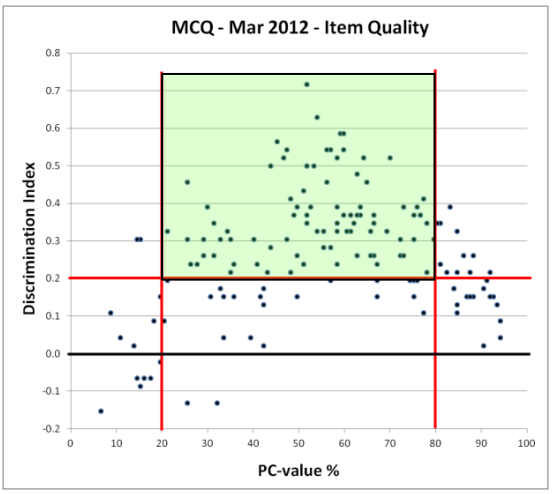


Figure 5.1(a): MCQ - March 2012

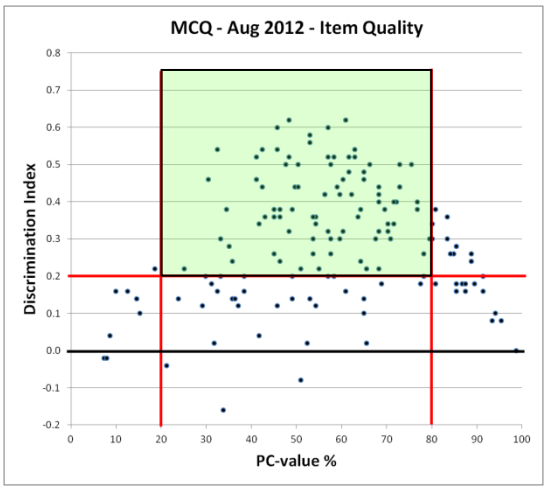


Figure 5.1(b): MCQ - August 2012

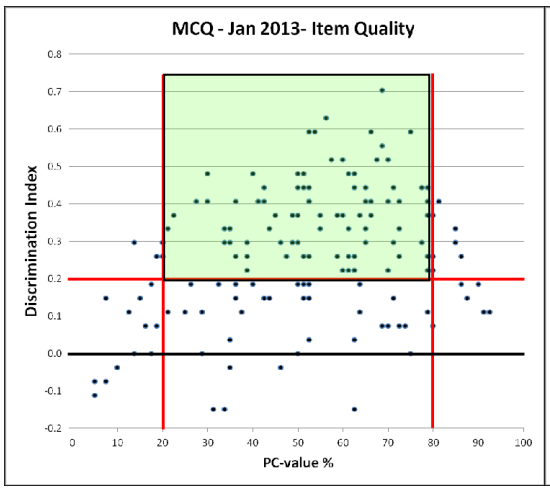


Figure 5.1(c): MCQ - January 2013

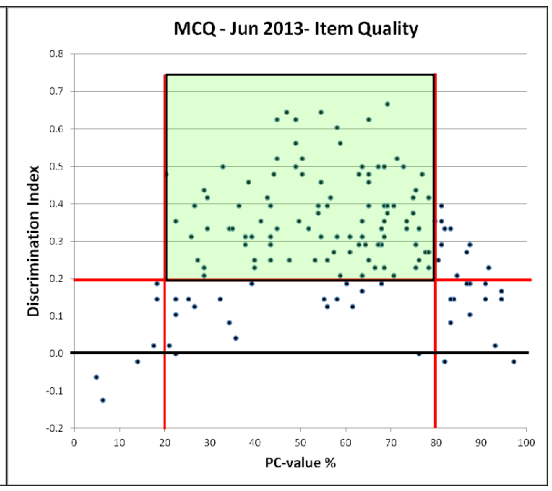


Figure 5.1(d): MCQ – June 2013

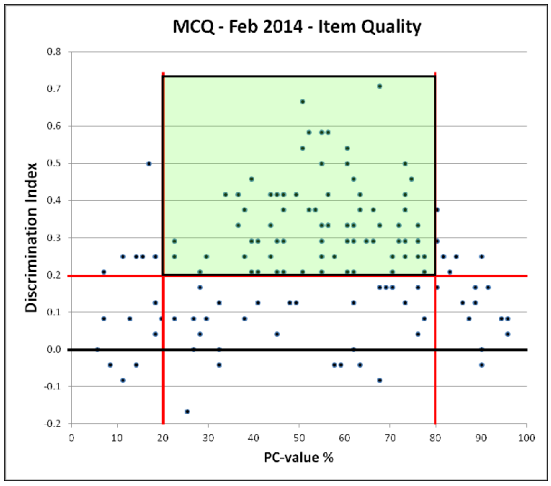


Figure 5.1(e): MCQ – February 2014

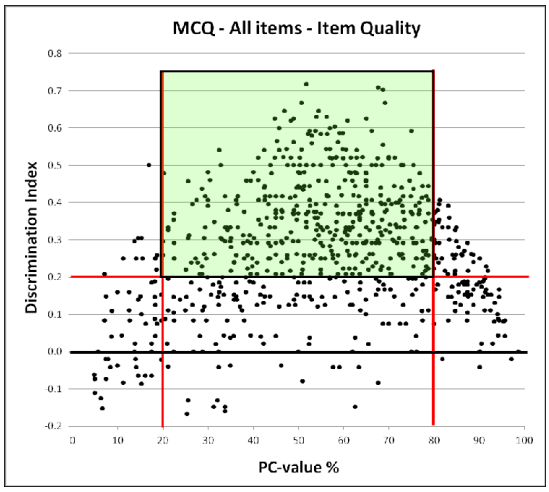


Figure 5.1(f): MCQ All items (n=750)

5.2.1.2 *MCQ test reliability analysis - all 150 items*

Cronbach's alpha coefficients for the five MCQ papers included in the study ranged from 0.89 – 0.93, with a median value of 0.92. The *SEM* of the five test cycles ranged from 4.2 - 4.3, with a median of 4.2.

5.2.2 MCQ test performance data - the 30 tracker items

Within each of the five 150-item MCQ papers included in the study there were 30 questions repeated in each test cycle. These tracker questions, treated as a 'mini-test' within each test, made it possible to compare the performance of consecutive cohorts of candidates. Owing to a technical error the January 2013 paper included only 26 of the 30 tracker items. The 'mini-test' scores for this paper were converted to a mark out of 30 to enable comparison with the other papers.

Descriptive statistics of the tracker items (from Table 5.3)
The *maximum* scores for the 30-item mini-test contained in each of the five 150-item MCQ papers, ranged from 88 - 100% (range= 12%) with a median score of 92%. The *minimum* scores ranged from -10 – 6% (range= 16%), with a median score of 3%.

An ANOVA analysis showed no statistically significant differences (p=0.548) between the *mean scores* of the five MCQ mini-tests. The mean scores ranged from 47.2 – 51.1%, with a median score of 48.7%. The *standard deviation* around the mean scores ranged from 18.8 – 23.6%, with a median value of 21.9%. The *95% confidence interval* (CI) for the mean score of each mini-test is provided in Table 5.3. The width of the CIs for the five mini-tests ranged from 7.3 – 9.0%, with a median value of 7.9%.

TABLE 5.3: THE PART I MCQ TEST - TRACKER PERFORMANCE DATA (30 ITEMS)

Part I MCQ - Tracker items (30)		Cycles (n= candidates)					Analysis of 5 cycles			
		Mar2012	Aug 2012	Jan2013*	Jun2013	Feb2014				
Analysis Component	**Analysis Descriptor**	n= 137	n= 151	n= 80	n= 143	n= 71	**Min**	**Max**	**Range**	**Median**
Descriptive statistics	Maximum score (%)	100	93	88	92	88	88	100	13	92
	Minimum score (%)	-10	-3	-6	6	5	-10	6	16	-3
	Mean score (%)	47.2	51.1	47.9	50.4	48.7	47.2	51.1	3.9	48.7
	Standard Deviation (SD) (%)	23.6	22.5	20.3	21.9	18.8	18.8	23.6	4.8	21.9
	95% confidence intervals (CI) of mean (%)	43.2 - 51.1	47.4 - 54.7	43.4 - 52.4	46.8 - 54.1	44.3 - 53.2	-	-	-	-
	95% CI width (%)	7.9	7.3	9.0	7.3	8.9	7.3	9.0	1.7	7.9
Item analysis	Mean Item Difficulty (PC -value)	0.56	0.59	0.56	0.58	0.57	0.56	0.59	0.02	0.57
	Mean Item Discrimination Index (DI)	0.41	0.39	0.32	0.38	0.29	0.29	0.41	0.11	0.38
	Median Item Discrimination Index (DI)	0.37	0.41	0.37	0.37	0.31	0.31	0.41	0.10	0.37
	**Item Quality Index (IQI) (%)	97	83	81	90	67	67.0	97.0	30.0	83.0
	Correlation with whole paper (150 items) r=	*0.93*	*0.93*	*0.86*	*0.91*	*0.89*	0.86	0.93	0.07	0.91
	Correlation with whole paper (150 items) r^2=	0.86	0.86	0.74	0.83	0.78	0.74	0.86	0.12	0.83
Test reliability analysis	Cronbach's alpha coefficient	0.82	0.82	0.72	0.81	0.73	0.72	0.82	0.10	0.81
	Standard Error of Measurement (%)	9.93	9.63	10.82	9.62	9.71	9.6	10.8	1.2	9.7

* The January 2013 paper only had 26 tracker items

** See text for explanation of IQI

Red italic numbers: All correlations were statistically significant (p <0.001)

5.2.2.1 *MCQ tracker item analysis*

The *mean PC-value* of the tracker items for the five mini-tests ranged from 0.56 - 0.59 with a range of 0.02 (due to rounding) and a median value of 0.57. An ANOVA analysis reported no statistically significant difference (p=0.959) between the mean PC-value of the five mini-tests (*cf.* Table 5.3).

The *mean DI* of the tracker items for each of the five mini-tests ranged between 0.29 – 0.41, with a median of 0.38. A Welch ANOVA analysis showed a statistically significant difference between one or more pairs of DI mean scores across the five mini-tests (p=0.015). The post-hoc Games-Howell test showed a statistically significant difference between the March 2012 and February 2014 papers (p=0.030). These variants of the standard one way ANOVA omnibus test and post-hoc test were used, because there was a statistically significant difference in the homogeneity of variance between the DI values of the five mini-tests (Levene's test).

Although the tracker items' DI distributions per paper were normally distributed by Normal Q-Q plot analysis, the medians are also provided in Table 5.3. The January 2013 cohort, which was a small cohort (n=80), had a mean DI of 0.32 and a median of 0.37 across the 26 items contained in its mini-test.

Since the same items were used in all the papers, the average performance *per item* was calculated over the 5 cycles. These 'within item' mean PC-values % and DIs were used to construct the Item Quality plot for the tracker items – *cf.* Figure 5.2 below. The mean mini-test PC-value, expressed as a percentage, and the average DI were 57% and 0.36, respectively.

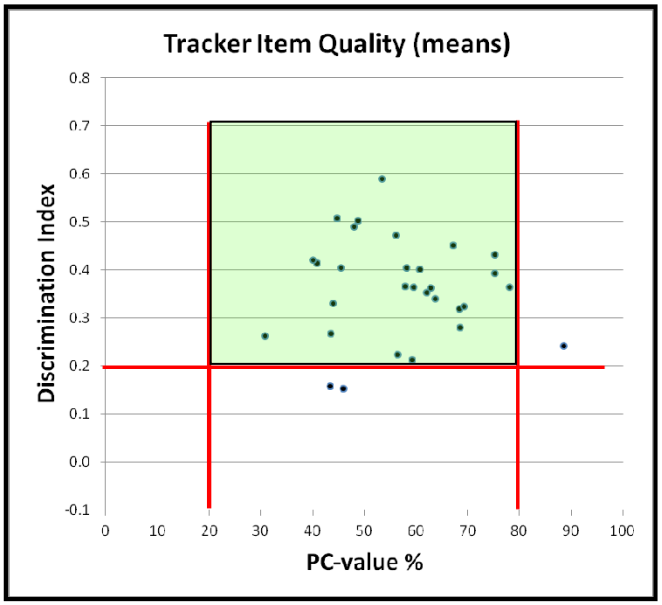


FIGURE 5.2: ITEM QUALITY PLOT FOR THE TRACKER MCQ ITEMS (MEANS)

The *item quality index* (IQI) of the 30 tracker MCQ items, as calculated from Figure 5.2, was 90% (percentage of items in the green zone). The IQI for the tracker items in each of the mini-tests ranged from 67 – 97%, with a median of 83% (*cf.* Table 5.3).

The *Pearson correlations* (r) between the candidates' performances on the tracker mini-test items and the whole 150-item MCQ test ranged from 0.86 to 0.93 and were statistically significant in all cases (p<0.001). The corresponding r^2 values ranged between 0.74 – 0.86, with a median value of 0.83.

5.2.2.2 *MCQ tracker item reliability analysis*

Cronbach's alpha coefficient for each of the five mini-tests ranged from 0.72 – 0.82, with a median value of 0.81. The smaller cohorts (January 2013 and February 2014) had lower alpha coefficients (0.72 and 0.73), as compared to the three larger cohorts (0.81 - 0.82).

The *SEM* of the five mini-tests ranged from 9.6% – 10.8%, with a median value of 9.7%.

5.2.3 MCQ test outcome using the Angoff method

Angoff panel participants

There were 12-18 subject expert judges (CoP examiners) involved in each of the five Part I MCQ test Angoff standard setting procedures. On average the panellists had 9.9 years of experience as FCP examiners.

Pass marks

The *Angoff pass marks* generated by the judges for the five cycles of the whole 150-item Part I MCQ test ranged from 56 – 65% (9%), with a median score of 59% (*cf.* Table 5.4).

The Angoff pass marks for the five tracker MCQ mini-tests ranged of 55 – 81% (26%), with a median of 59% (*cf.* Table 5.5).

Failure rates

The *failure rates* produced by the Angoff pass marks of the *overall* MCQ tests' ranged from 64.9 - 97.5% (range= 32.6%), with a median failure rate across the five MCQ tests of 78.9% (*cf.* Table 5.4).

The failure rates, produced by the Angoff method, for the *tracker* mini-tests ranged from 57.7 - 96.3% (range= 38.5%), with a median failure rate across the five mini-tests of 71.5% (*cf.* Table 5.5).

The pass marks and resulting failure rates produced by the two different standard setting methods used in this study are illustrated in Figure 5.5(a-b) for the five 150-item MCQ tests and Figure 5.6(a-b) for the tracker mini-tests, respectively. The failure rates of the previous fixed 50% pass mark are also included in Figure 5.5b and Figure 5.6b.

TABLE 5.4: THE PART I MCQ TEST - STANDARD SETTING DATA (ALL ITEMS)

Part I MCQ exam - All items (150)		Cycles (n= candidates)					Analysis of 5 cycles			
		Mar2012	Aug 2012	Jan2013	Jun2013	Feb2014				
Analysis Component	**Analysis Descriptor**	n= 137	n= 151	n= 80	n= 143	n= 71	**Min**	**Max**	**Range**	**Median**
Standard setting analysis and resulting failure rates	95th percentile	71.0	74.0	63.0	77.0	63.5	63	77	14	71.0
	Cohen65 pass mark	**46**	**48**	**41**	**50**	**41**	41	50	9	46.0
	Angoff pass mark	**64**	**58**	**65**	**59**	**56**	56	65	9	59.0
	Cohen65 FAILURE RATE %	48.9	43.0	41.3	51.0	32.4	32.4	51.0	18.7	43.0
	Angoff FAILURE RATE %	89.1	64.9	97.5	65.7	78.9	64.9	97.5	32.6	78.9
	50% FAILURE RATE %	60.6	49.0	60.0	51.0	57.7	49.0	60.6	11.6	57.7
Angoff validity analysis	Pearson Correlation - Angoff % vs PC-value r=	0.12	*0.45*	*0.33*	*0.47*	*0.48*	0.12	0.48	0.36	0.45
	Correlation - Angoff % vs PC-value r^2=	0.01	0.21	0.11	0.22	0.23	0.01	0.23	0.22	0.21
Angoff method reliability analyses	SEMean Angoff pass mark	4.4	4.2	4.4	6.0	4.3	4.2	6.0	1.8	4.4
	Max SEMean allowed for reliable Angoff result[1]	2.1	2.1	2.2	2.1	2.1	2.1	2.2	0.0	2.1
	Standard deviation (SD) of Angoff pass marks	15.8	16.3	15.1	25.4	16.8	15.1	25.4	10.3	16.3
	Max Angoff SD allowed for reliable result[2]	3.7	3.9	3.5	4.0	3.2	3.2	4.0	0.8	3.7
	Inter-rater Reliability - Light's Kappa coefficient[3]	0.24	0.17	0.21	0.15	0.24	0.15	0.24	0.09	0.21
	Light's Kappa - 95% confidence intervals (CI)	0.21 - 0.27	0.15 - 0.19	0.18 - 0.26	0.14 - 0.17	0.22 - 0.26	-	-	-	-
	Light's Kappa - 95% CI width	0.06	0.04	0.08	0.03	0.04	0.03	0.08	0.05	0.04

SEMean = Standard Error of the Mean; *Red italic numbers: Correlations were statistically significant (p <0.001)*

[1] Cohen, Kane, Crooks 1999

[2] Meskauskas 1986

[3] Hallgren 2012

TABLE 5.5: THE PART I MCQ TEST - TRACKER STANDARD SETTING DATA (30 ITEMS)

Part I MCQ - Tracker items (30)		Cycles (n= candidates)					Analysis of 5 cycles			
		Mar2012	Aug 2012	Jan2013*	Jun2013	Feb2014				
Analysis Component	**Analysis Descriptor**	n= 137	n= 151	n= 80	n= 143	n= 71	**Min**	**Max**	**Range**	**Median**
Standard setting analysis and resulting failure rates	95th percentile	84.8	83.0	76.1	83.9	78.0	76.1	84.8	8.7	83.0
	Cohen65 pass mark	**55**	**54**	**49**	**55**	**51**	49.0	55.0	6.0	54.0
	Angoff pass mark (mean of judges)	**62**	**67**	**81**	**63**	**55**	55.0	81.0	26.0	63.0
	Cohen65 FAILURE RATE %	65.7	51.0	43.8	58.0	46.5	43.8	65.7	21.9	51.0
	Angoff FAILURE RATE %	73.0	71.5	96.3	67.8	57.7	57.7	96.3	38.5	71.5
	50% FAILURE RATE %	54.7	43.0	45.0	45.5	46.5	43.0	54.7	11.7	45.5
Angoff validity analysis	Pearson Correlation - Angoff % vs PC-value r=	-0.01	*0.41*	0.21	0.32	0.32	-0.01	0.41	0.42	0.32
	Correlation - Angoff % vs PC-value r^2=	0.00	0.17	0.04	0.10	0.10	0.00	0.17	0.17	0.10
Angoff method reliability analyses	Inter-rater Reliability - Light's Kappa coefficient[1]	0.26	0.18	0.18	0.13	0.24	0.13	0.26	0.13	0.18
	Light's Kappa - 95% confidence intervals (CI)	0.21 - 0.31	0.13 - 0.23	0.13 - 0.24	0.11 - 0.16	0.21 - 0.27	-	-	-	-
	Light's Kappa - 95% CI width	0.10	0.10	0.11	0.05	0.06	0.05	0.11	0.06	0.10

Red italic numbers: Correlations were statistically significant (p=0.025)

[1] Hallgren 2012

Validity of the Angoff method

A *validity-correlation plot* (*cf.* Chapter 2), was used as a quantitative measure of the internal validity of the Angoff procedure. This was done by determining the Pearson correlation coefficient (r), for the relationship between the predicted test item difficulty (Angoff rating expressed as percentage) and actual test item difficulty (PC-value expressed as a percentage) of the individual MCQ test items. The Pearson correlation values (r) for each of the five 150-item MCQ tests ranged from 0.12 to 0.48, with a median of 0.45. All the correlation coefficients were statistically significant ($p < 0.001$), except the 0.12 value ($p = 0.15$). The corresponding r^2 values ranged from 0.01 to 0.23 (cf. Table 5.4 or 5.6).

Figure 5.3 is the validity-correlation plot for all 750 MCQs used in the five test cycles. The Pearson correlation coefficient (r) was 0.37, which was statistically significant ($p < 0.001$) and the corresponding r^2 value was 0.14.

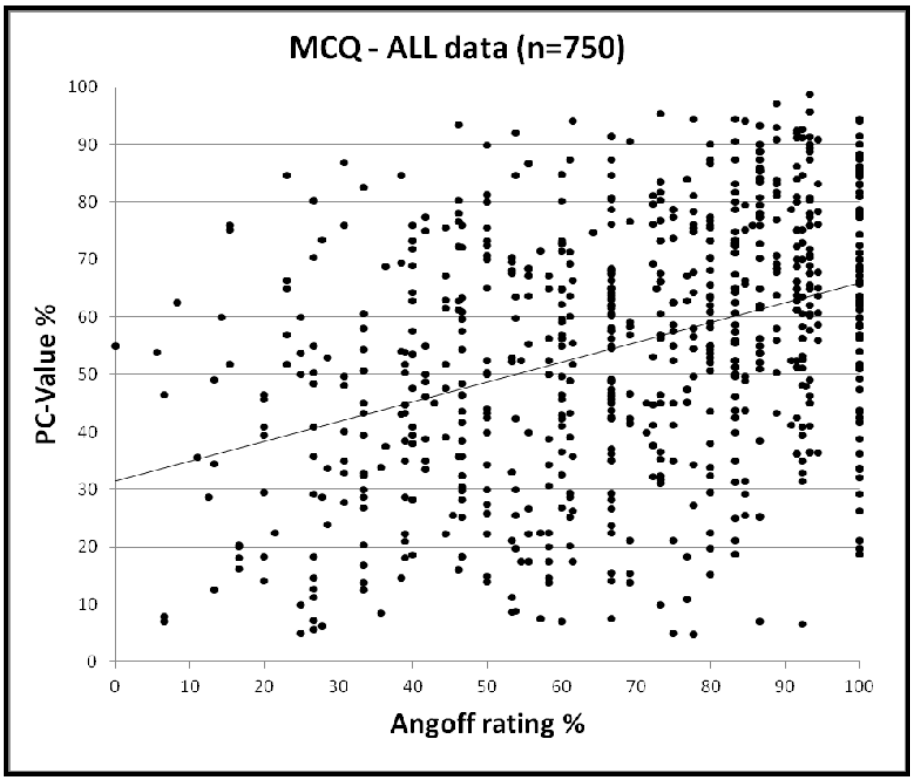


FIGURE 5.3: CORRELATION PLOT FOR ALL MCQ ITEMS (n=750)

The data for the *tracker items* were extracted and analysed separately. The Pearson correlation coefficients, provided in Tables 5.5 and 5.6, ranged from -0.01 to 0.41, with a median value or 0.32 for the five mini-tests studied and the corresponding r^2 values were between 0.00 and 0.17. Only the 0.41 correlation from the August 2012 mini-test data was statistically significant (p=0.025).

The validity-correlation between the PC-values, expressed as a percentage, and the mean Angoff rating, expressed as a percentage, for each tracker item included in the five tests studied, is shown in Figure 5.4 below. The Pearson correlation coefficient (r) was 0.38, which was statistically significant (p =0.037) and the corresponding r^2 was 0.15.

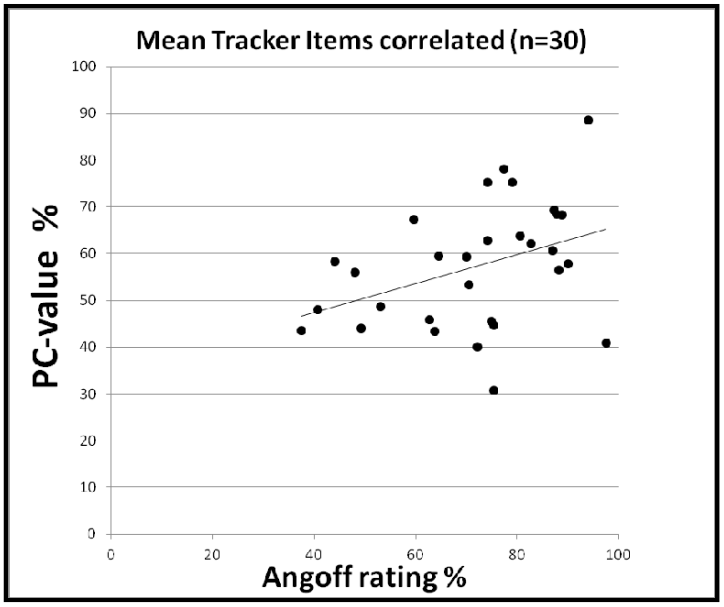


FIGURE 5.4: VALIDITY-CORRELATION PLOT FOR THE MEAN TRACKER MCQ ITEMS (n=30)

Table 5.6 summarises the Angoff data from the five full 150-item MCQ tests and the corresponding 30-item tracker mini-test in terms of four indicators:

1. The overall Angoff pass mark per test;
2. The mean Angoff rating %;
3. The mean PC-value %;
4. The Pearson correlation coefficient (r) between 2 and 3.

The reason for the difference between the Angoff pass mark % and the Angoff rating % is due to the negative marking system used by the CoP.

Table 5.6 revealed that the mean Angoff rating %, which is the predicted performance of a *borderline* candidate, and the resulting Angoff pass mark % were consistently *higher* in every cycle than the mean PC-value, which is an indication of the performance of the *average* candidate. This pattern was evident in both the full MCQS tests as well as mini-tests.

TABLE 5.6: THE PART I MCQ TEST: ANGOFF ANALYSIS SUMMARY

Cycle (n)	Angoff pass mark* (%)		Mean values				Pearson correlation coefficient (r)	
			Angoff rating (%)		PC-value (%)			
	Full	Tracker	Full	Tracker	Full	Tracker	Full	Tracker
Mar 2012 (137)	64	62	71.2	69.6	55.5	56.3	0.12	-0.01
Aug 2012 (151)	58	67	66.5	73.9	56.8	58.6	*0.45***	*0.41*#*
Jan 2013 (80)	65	81	72.0	84.5	51.7	56.3	*0.33***	0.21
Jun 2013 (143)	59	63	66.8	70.2	57.4	58.3	*0.47***	0.32
Feb 2014 (71)	56	55	65.0	63.7	53.5	56.9	*0.48***	0.32
Median	**59**	**63**	**66.8**	**70.2**	**55.5**	**56.9**	**0.45**	**0.32**
Range	**9**	**26**	**7.0**	**20.8**	**5.7**	**2.3**	**0.36**	**0.42**

* Negative marking applied

*Red italic numbers: Correlations were statistically significant, p-values <0.001*** and p=0.025#

Reliability of the Angoff method

Three indicators of reliability were used and triangulated in this study to form a judgement regarding the reliability of the Angoff process used for the MCQ test cycles. The data from these three measures are provided in Table 5.4.

The first method used, described by Cohen *et al.* (1999), showed that none of the five cycles had a reliable outcome (highlighted in red - *cf.* Table 5.4).

The second method, described by Meskauskas (1986), showed that none of the five cycles had a reliable outcome (highlighted in red – *cf.* Table 5.4).

The third, and probably the most robust reliability estimation method used in this study, was the *inter-rater reliability* (IRR) calculation. Light's Kappa coefficient method, a specific form of IRR calculation, was used because the Angoff method generated binary rating data by more than two raters (*cf.* Hallgren 2012:5-6). The findings from this method are also reported in Table 5.4. The Kappa coefficients for the five test cycles were between 0.15 – 0.24, with narrow 95% confidence intervals (CIs) of between 0.03 – 0.08 per test.

The reliability of the five Angoff procedures for the *tracker mini-test items* was only calculated with Light's Kappa IRR measurement. The Kappa coefficients of the five Angoff processes on the mini-test items in each cycle are shown in Table 5.5. They ranged from 0.13 – 0.26, with 95% CIs of between 0.05 – 0.11 per mini-test (*cf.* Table 5.5)

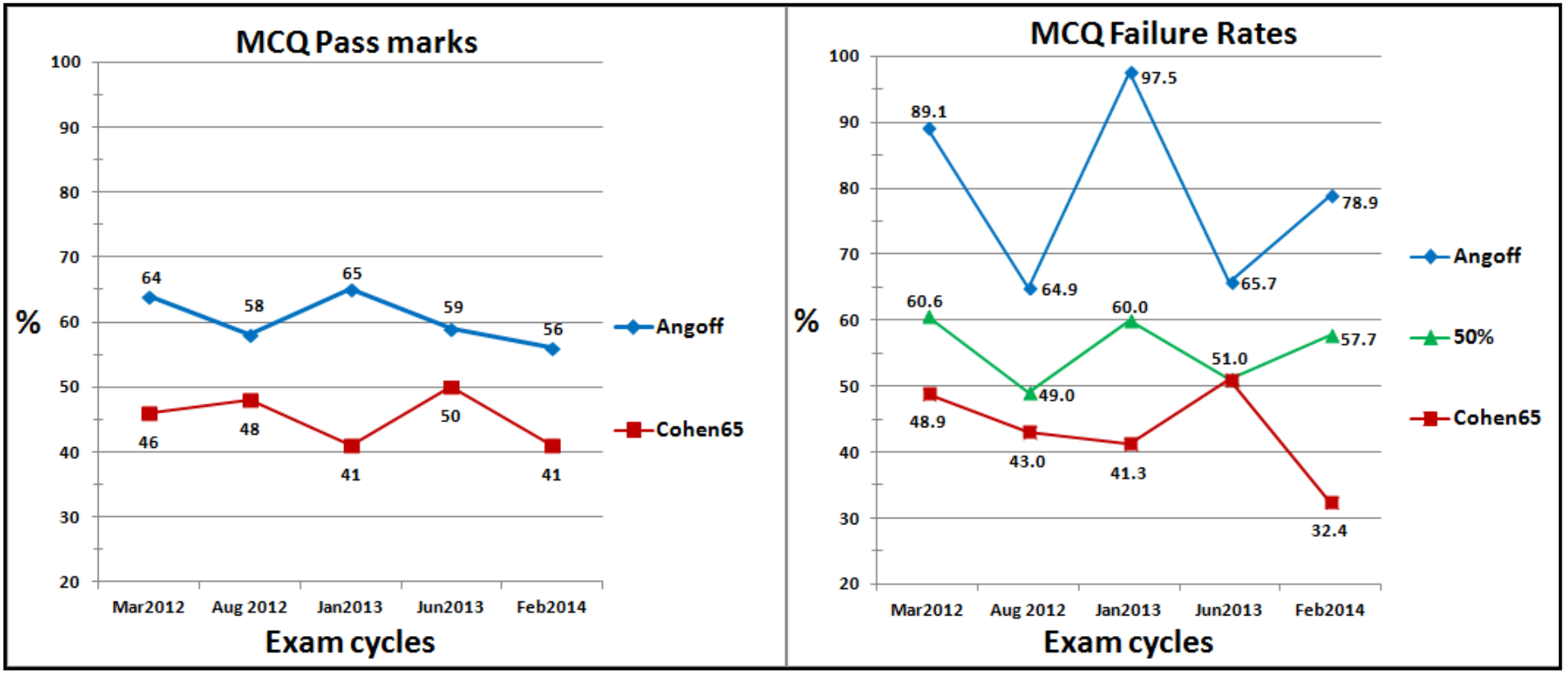


FIGURE 5.5(a-b): MCQ (FULL TEST) PASS MARKS (a) AND RESULTING FAILURE RATES (b)

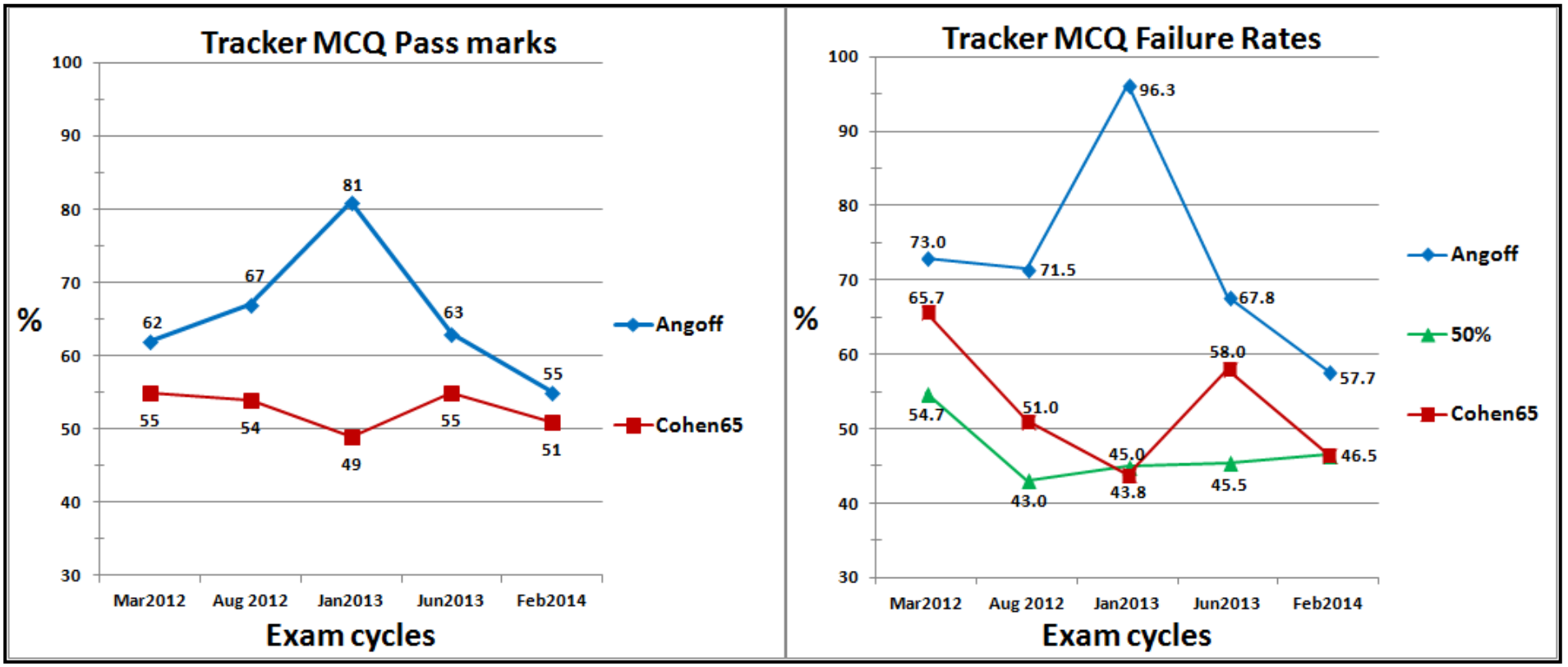


FIGURE 5.6(a-b): MCQ (TRACKER MINI-TEST) PASS MARKS (a) AND RESULTING FAILURE RATES (b)

5.2.4 MCQ test outcome using the Cohen method

The data derived from the five sets of MCQ test data using the Cohen method of standard setting are shown in Table 5.4. The *Cohen65 pass mark* for the each of the five cohorts ranged from 41 – 50% (range= 9%), with a median score of 46%.

The resulting failure rates were between 32.4 – 51% (range= 18.7%), with a median failure rate of 43%. The two smaller cohorts, January 2013 (n=80) and February 2014 (n=71), both had Cohen65 pass marks of 41%, whereas the larger three cohorts had pass marks between 46 – 50% (*cf.* Table 5.4).

The same pattern emerged from the tracker mini-test data (*cf.* Table 5.5). The *Cohen65 pass marks* for the five mini-tests ranged from 49 – 55% (range= 6%), with a median score of 54% and resulting failure rates ranged between 43.8 – 65.7% (range= 21.9%), with a median of 51%. The two smaller cohorts, January 2013 (n=80) and February 2014 (n=71), had Cohen65 pass marks of 49 - 51%, whereas the larger three cohorts' pass marks ranged from 54 – 55% (*cf.* Table 5.5).

The pass marks and resulting failure rates produced by the two different standard setting methods used in this study are illustrated in Figure 5.5(a-b) for the five 150-item MCQ tests and Figure 5.6(a-b) for the tracker mini-tests, respectively. The failure rates of the previous fixed 50% pass mark are also included in Figure 5.5b and Figure 5.6b.

The *95th percentiles*, from which the Cohen65 pass marks were derived, ranged from 63 – 77%, with a median value of 71% (*cf.* Table 5.4). Again, the 95th percentiles of the two smaller two cohorts were similar (63%), but lower than the larger three cohorts (71-77%).

The 95th percentiles for the tracker test items for the three larger cohorts were between 83 – 84.8% and for the smaller two cohorts between 76.1 – 78.0%.

5.2.5 MCQ test outcome using a 50% pass mark

The previous traditional *fixed 50% pass mark* practice would have resulted in failure rates between 49.0 – 60.6% (range= 11.6%), with a median failure rate of 57.7% across the five 150-item MCQ tests (*cf.* Table 5.4). The fixed 50% pass mark failure rates for the mini-tests, over the five cycles, ranged between 43.0 – 54.7% (range= 11.7%), with a median failure rate of 45.5% (*cf.* Table 5.5).

5.3 THE PART I MCQ TEST DISCUSSION

This chapter contributes data towards answering Research Question 3: "Is using the Cohen method, as compared to Angoff method, a more appropriate way of determining the pass mark for the written assessments of the FCP (SA) examination (Part I MCQ test, Part II Objective test and Part II SEQ test)?"

In this chapter the results of the second component of the study, relating specifically to evaluation of the performance of the Angoff and Cohen methods of standard setting, using the entry-level FCP (SA) Part I *MCQ test* performance data, have been reported, and will now be compared during the discussion of their respective outcomes.

The discussion is structured broadly in the same order in which the results were presented in the preceding text.

5.3.1 Candidates and Cohorts

This study included the results of 582 candidates who sat the MCQ test (five test cycles) between March 2012 and February 2014. This is a good dataset size to study for five test cycles, but as McManus *et al.* (2014:15) shows, some findings from initial data needs longer time periods and more data to be effectively studied, especially the introduction of different standard setting methods. This finding from the literature was acknowledged during the analysis and review of the results and outcomes of this study.

As can be seen from the data, the number of candidates sitting the FCP (SA) Part I MCQ test is increasing. This is most likely due to the HPCSA ruling in 2010, which

established a single, national postgraduate licensing process (HPCSA 2010). Although local universities could still offer *entry*-level examinations, this practice was phased out completely by 2012. Therefore since 2013, all candidates wishing to pursue specialist physician training have to undertake the FCP (SA) Part I MCQ test. Unfortunately, unlike for the Part II examination, there were no published historical candidate data available to compare the numbers of candidates sitting the Part I examination.

The Part I MCQ test cohort sizes, included in this study, showed marked variability in numbers (*cf.* Table 5.2). The five cohort sizes ranged from 71 – 151, with a mean size of 116, and a median size of 137. The reason for the wide range and seemingly large difference between the mean and the median as noted in Table 5.2, where the cohort sizes across the five cycles revealed three larger cohorts - March 2012 (n=137), August 2012 (n=151) and June 2013 (n=143) as well as two distinctly smaller cohorts – January 2013 (n=80) and February 2014 (n=71).

A possible reason for this variation in cohort size might be due to a change in the *timing* of the Part I examination cycles. Before 2013, the FCP (SA) Part I examination was offered in March and August annually, but from 2013 the CMSA brought the dates forward to January/beginning of February and June each year. The effect was that considerably fewer candidates opted to write the Part I in the first cycle of 2013. Possible reasons for the change were perhaps due to the South African medical education graduation and 'change-over/move-on' dates usually occurring in December every year, since the system is organised in calendar years.

Whatever the cause of the cohort size variation was, it did seem to have an effect, as explained later in the discussion. Candidates' number of attempts, academic history and demographic information were not readily available to include in this study and therefore, it was not possible to establish if the different cohorts were comparable, in terms of these candidate variables. This information would have provided helpful insight, as reported in other studies (McManus *et al.* 2014:2-19; Wakeford, Denney, Ludka-Stempien, Dacre & McManus 2015:2-12), into the relative heterogeneous nature of the Part I cohorts, as explained in Chapter 3 (section 3.5.4).

The hypothesis of the study was, in terms of the Cohen method, that there would be a sufficient number of top performing candidates in every cohort and that fluctuations in the cohort size would not be a real factor in determining the pass mark. Given the

variations in the cohort sizes noted in this study, the impact of the different cohort sizes on the Cohen pass mark, became an important outcome in this study in terms of the utility of the Cohen method (robustness). The outcomes of the three larger cohorts were compared with the two smaller ones, but the small sample size (five cohorts in total) for these comparisons is recognised and more data over a longer time are needed to come to a more definitive conclusion about the effects of cohort size and composition on the Cohen method.

5.3.2 MCQ test candidate performance data

5.3.2.1 *Performance data on all 150 MCQ items*

The effect of *negative marking* on the MCQ test was clearly evident in the performance data for the five cohorts, especially its effect on the poorly performing candidates. Over the five cycles, maximum scores had a range of a mere 4%, between 78 – 82% (median of 79%) compared with the minimum score range of 20%, with scores ranging from -2% to 18% and a median of 14% (*cf.* Table 5.2). It was concerning to see a candidate with a negative minimum score in the March 2012 cohort, which reflects the extent of examination underpreparedness in the MCQ test.

As mentioned, the effect of negative marking on the mean scores was also noted, by increasing the variance between the five mean scores. An ANOVA analysis showed that there were statistically significant differences between the means scores, with a post-hoc test indicating the difference was between the January 2013 cohort and the August 2012 (p=0.038) and June 2013 (p=0.018) cohorts, respectively. Interestingly, when negative marking was not used, the mean scores of the five cohorts showed *no* difference (p=0.959). This was established by an ANOVA analysis using the mean PC-values for each cohort, which does not include negative marking. This was evidence that *Construct-irrelevant variance* (CIV) was added to the test scores by using negative marking, as was predicted by the literature (Downing 2003a:671). A discussion on negative marking and CIV was provided in Chapter 2.

This meant that the performance of the five cohorts was similar for the five MCQ tests, once negative marking was removed. Therefore, it can be reasonably deducted that the *mean* academic ability of the five cohorts of candidates were comparable and the

difficulty of the MCQ test papers were also comparable. Analysis of the tracker MCQ mini-test provided further evidence in support of this finding (see below).

The spread of the candidates' scores were reasonably similar across the five cycles as indicated by the 3% difference in the SDs of the five cohorts, which ranged from 12.9 – 15.9% (*cf.* Table 5.2). Interestingly, the smaller two cohorts (January 2013, February 2014) had the lowest SDs and hence the narrowest distributions of scores. This could possibly have affected the 95th percentile scores in these cohorts and the resultant Cohen65 pass marks.

Item analysis

As discussed previously, the mean PC-values across the five MCQ papers showed no difference and ranged from 0.52 – 0.57, which suggests a moderate mean test difficulty (Sim & Rasiah 2006:69). If the PC-values are converted to percentage scores, they would be the mean 'number correct' % scores of the MCQ papers. These PC-value% scores were used in the correlation analyses.

The mean Discrimination Index (DI) values ranged from 0.23 – 0.30, with a median value of 0.27 for the five MCQ tests, which were above the acceptable/good level of 0.20 suggested in the literature (Downing 2009a:L4747) (*cf.* section 2.2.4.3).

The item quality index (IQI), expressed as a percentage value, ranged from 55 - 69%, with a median of 59%, which indicated that at least 55% of the items in each MCQ paper had an acceptable combination of difficulty (PC-value%) and discriminatory ability (DI). If all 750 MCQ test items were combined, the overall IQI was 61%. This was a new composite item analysis indicator developed for this study and as such there were few references to it in the literature. However, a recent small undergraduate dental study in Pakistan reported an IQI of 64% (using slightly different PC-values and DI quality ranges) in their 50-item, best-of-five, MCQ paper (Hingorjo & Jaleel 2012:142). Most references to the literature merely states what constitutes acceptable quality of the individual components (e.g. PC-value and DI) and most studies then report their findings to those criteria, such as Ware & Vik (2009:240-241). Logically, the higher the IQI of a paper the better, but what constitutes an acceptable minimum IQI for a paper has not yet been defined in the literature. However, a reasonable IQI is probably one where more than 50% of the items in the test are of

good quality, since some authors have argued that tests should contain more than 50% items satisfying individual quality indicators, such as DI (Ware & Vik 2009:241).

Reliability analysis

The MCQ papers included in this study had excellent reliability, as compared to the literature guidelines for high-stakes tests (Tighe *et al.* 2010:2) (see section 2.2.4.4). Cronbach's alpha coefficients were all 0.89 or above and the SEMs 4.3% or less. The reasons for the similar reliability credentials (Cronbach's alpha coefficients and SEMs) of the five MCQ papers reflect adequate sampling, together with similar test difficulty and candidate performance, which produced similar score distribution (standard deviations - SD) for the papers. The SEM range is very narrow since it is robust for the SD variation in each paper (Tighe *et al.* 2010:2), and therefore it is fair to report that the MCQ papers had essentially the same reliability.

5.3.2.2 *Performance data on the 30 tracker MCQ items*

As explained in Chapter 3 (see section 3.5.2.1), the 30 tracker MCQ items, which constituted the repeating mini-tests contained inside the five MCQ tests, were selected based on their good psychometric properties and their representativeness of the full MCQ papers (blueprint). The performance data on the full 150-item MCQ tests indicated the same mean academic ability of the five cohorts (no difference in the mean PC-values) and similar distributions of their scores (SDs of the cohorts ranged from 13 - 16%). The five cohort performances on the mini-test reflected a combination of the tracker item selection criteria, smaller sample size of test items and the mean academic ability of the cohorts.

The cohorts' maximum and minimum scores on the mini-test were higher and lower respectively than on the full MCQ papers. Mini-test maximum scores ranged from 88 - 100% (12%), with a median score of 92% and minimum scores -10 – 6% (16%), with a median score of -3% (due to negative marking). There were no statistically significant differences reported with an ANOVA analysis between the mean scores of the five cohorts on the mini-test (p=0.548) or between their respective mean PC-values (p=0.959). This suggests that the CIV introduced to the scores by negative marking did not have a significant effect on the mini-test scores, most likely due to the smaller sample size of the mini-test.

The median of the five mini-tests' mean scores was 48.7% and the median of the five full MCQ tests' mean scores was 46.2%, a difference of only 2.5%. The median of the five mini-tests' mean PC-values were 0.57 and the median of the five full MCQ tests' mean PC-value was 0.55 a difference on only 0.02 or 2%. This suggests that the mini-test and the full MCQ test were of similar difficulty. However, the standard deviations (SD) of the five cohorts on the mini-test were much wider (range 18.8 – 23.6%) compared to the full MCQ paper (range 12.9 – 15.9%). The same trend was noted with the width of 95% confidence intervals of the means on the mini-test were wider (range 7.3 – 9.0%) compared to the full MCQ paper (range 4.9 – 6.3%). This was probably the effect of the higher mean DIs of the mini-tests (ranged from 0.29 – 0.41) and their lower reliability (Cronbach's alpha range 0.72 – 0.82, SEM range 9.6 - 10.8%) due to a much smaller sample size, compared to the full MCQ tests. Interestingly, similar to the full MCQ tests, the smaller two cohorts again had the lowest SDs in the mini-test and hence, the narrowest distributions of mini-test scores.

Tracker Item analysis

As mentioned previously, the mean PC value reflects the mean difficulty of the test items and for the mini-tests there was no statistically significant difference between the cohorts. The mean performance on the tracker items was the same in the five cycles of data reviewed.

The mean DIs of the cohorts, ranging from 0.29 – 0.41, showed that they all discriminated effectively between top and bottom performing candidates, and the mini-test DI median of 0.38 was considerably higher than the full MCQ test DI median of 0.27. This meant the tracker items performed as expected, since they were selected for their high DI potential. An interesting finding that emerged was that although the tracker items had the same mean difficulty for all cohorts, they did not have the same mean discrimination power for all cohorts. The cohort size seemed to play a role in the mean DI, with the three larger cohorts ranging between 0.38 - 0.41 and the two smaller cohorts ranging from 0.29 - 0.32, which suggest the difference between the top and bottom 33% were similar *within* the large or small cohorts, but not *between* them, where there was at least a 6% difference. An ANOVA analysis across the five cohorts showed a statistically significant difference between the March 2012 and February 2014 cohorts (p=0.030) and not surprisingly March 2012 was a large cohort and February 2014 was a smaller one.

The mean IQI across the five cohorts of the 30 tracker items was 90%, which was encouraging given their intended purpose of forming a high quality mini-test to enable credible comparison of the candidates and the tests.

Another important characteristic of an effective set of test equating items, such as the tracker items, is the extent to which these items represented the full MCQ test. In this study the Pearson correlation coefficient (r), for each of the test cycles, confirmed that the 30-item mini-tests were representative of the 150-item full MCQ test. Disattenuated correlations were not needed given the strongly positive Pearson correlation coefficients. Therefore it can be deduced that the candidate performance in the tracker items were strongly representative of the candidate performances in the larger 150-item tests. This meant that the tracker items were a valid sample of the greater tests and hence, the deductions from the tracker data are representative of the larger papers.

Reliability analysis

As expected, the alphas coefficients were lower for the mini-tests than the full 150-item MCQ tests, due to the reduced sampling in the mini-tests, however the reduced reliability would have been counteracted to some extent by the wider score distributions for the mini-tests. This effect was clearly evident with the good reliability reported in the three larger cohorts (which had the highest SDs) with Cronbach's alpha coefficients of 0.81 – 0.82 and the weaker alpha coefficients noted for the two smaller cohorts (which had the lowest SDs) of 0.72 – 0.73.

This point is further emphasized by the other measure that is sensitive to width of distribution, the DI. The rank order of the SDs and mean DIs across the five cohorts were the same. This suggests that the performance data in the smaller cohorts were more densely distributed around a similar mean, across the five cohorts. As noted before, this could have implications for the Cohen method of standard setting, which is sensitive to the number and distribution of top-performing candidates at the top-end of the performance scale, in a given test.

Also, the DIs of the January 2013 test had a 5% lower mean (0.32) than median (0.37) and this meant that more tracker items had lower DIs (smaller difference between top and bottom 33% of candidates) in this cohort, dragging the mean lower, and thus fits the link between SD and DI. The same pattern was evident, but to a

lesser extent, in the other small cohort, February 2014, which had the narrowest SD of 18.8% (mean DI= 0.29 and median DI= 0.31), but the opposite was observed in the large March 2012 cohort, which had the widest SD of 23.6% (mean DI= 0.41 and median DI= 0.37). It seems the SD has an effect on the DIs, as suggested by the findings of this study.

Given the point above, the loss of four items in the January 2013 mini-test probably had a minimal effect on the Cronbach's alpha coefficient, but it did appear to have a clear effect on the SEM. The SEM values ranged from 9.62 - 9.93% across the four 30-item mini-tests, and 10.82 for the remaining *26-item* mini-test (January 2013). The omission of four test items in the January 2013 mini-test had a negative impact on the SEM (increased to 10.82%), as compared with the stable SEM values of 9.62 – 9.93% for the other 30-item mini-tests. SEM is a *test-centred* reliability measure, which is robust against variations in the performance of candidates on the test and hence, a reduction of four items decreased the test's reliability to an SEM of 10.82%. This emphasizes the importance of interpreting the SEM in conjunction with the alpha coefficient and the SD of test scores.

Therefore, in summary, the tracker items were of high quality, similar difficulty and strongly representative of the larger 150-item tests across all five cohorts. The different cohort sizes seem to have affected the SDs and DIs of the performance data. Despite the high quality of the mini-tests, their much smaller sample size compared to full MCQ test had a clear weakening effect on their reliability and emphasized the critical importance of a large sample of test items in high-stakes tests.

5.3.3 MCQ test Angoff data

The number of expert judges involved in the Yes/No Angoff procedures (12-18) was adequate according to the literature (Brandon 2004:68). These expert judges had a mean duration of involvement in the FCP (SA) examinations of about 10 years. They set Angoff pass marks ranging from 56 – 65% (range = 9%), with a median of 59%. These Angoff pass marks incorporated the negative marking system used in the Part I MCQ test, meaning that "no" rated items were scored negatively (-0.25). The Angoff pass marks, if used by the CoP, would have resulted in very high failure rates ranging from 64.9% to 97.5% (range= 32.6%), with median failure rate of 78.9%.

The Angoff pass marks generated by the judges for the tracker mini-tests, when their Angoff data was extracted from each cohort, was expected to be fairly similar for the five cohorts, since it's the *same* test and their mean performance and mean difficulty was the same for all the cohorts. However, this was not the finding. The mini-tests' Angoff pass marks showed considerable variation, ranging from 55 – 81% (range= 26%) which was greater than the Angoff pass marks generated for the full MCQ test. The resulting failure rates that these pass marks would have caused ranged from 57.7 - 96.3% (range= 38.5%). This raised serious concerns about the validity and reliability of the Angoff standard setting process as used in this study.

Angoff versus previous 50% pass mark practice

In contrast with the previous traditional fixed 50% pass mark practice, which is not linked to test difficulty and hence is regarded as indefensible (Van der Vleuten 2010:175), the Angoff method, as employed in the CoP, did worse. The 50% fixed pass mark would have resulted in failure rates of between 49 – 60.6% (range = 11.6%), with a median failure rate of 57.7% in the full MCQ test and failure rates of between 43.0 – 54.7% (range= 11.7%), with a median failure rate of 45.5% in the tracker mini-tests. These might still be unrealistically high failure rates, but they are lower than the Angoff failure rates to the extent that they do not even overlap (*cf.* Figure 5.6 a-b).

Angoff validity analysis

As explained in the literature review in Chapter 2, the Angoff rating represents the predicted difficulty of a test item for a borderline candidate and the PC-value is an indication of the actual test item difficulty. Although the absolute values between the predicted and actual difficulty of a test item cannot be compared, their relationship to each other should be linear and positive, since they are both indicators of the difficulty of a test item.

In the context of this explanation, a Pearson correlation coefficient of 0.37 (and r^2 of 0.14) for all 750 MCQ items, which was statistically significant (p <0.001) (*cf.* Figure 5.3), suggests that the Angoff judges were unable to tell the difference between a difficult item and an easy item for the cohort of candidates in question. Similar Pearson correlations (r) were observed for the five individual 150-item MCQ papers with r-values ranging from 0.12 to 0.48 and corresponding r^2 values from 0.01 to 0.23 (*cf.* Table 5.4 and 5.6). Furthermore, while there was no difference between the mean

PC-values across the five cycles of MCQ examination data, the mean Angoff rating and the corresponding Angoff pass mark for each test cycle varied greatly (*cf.* Table 5.6). This finding suggests that the Angoff method, as used in this study, was insensitive to test difficulty and, therefore, performing inappropriately.

The tracker item correlations between Angoff rating percentage and PC-value percentage had a range of 0.42 (-0.01 to 0.41) and this was on the *same* test with the PC-values staying almost the same - no difference in the means across the five cycles and range between means were 2.3% (*cf.* Table 5.6). The r^2 values of these five cycles of mini-tests meant that between 0 - 17% of the relationship between the predicted and actual difficulty of the items can be explained by the data. The rest is error or noise. This was additional evidence that the Angoff method, as used in the CoP, was not fit for purpose.

Furthermore, the mean Angoff ratings, the predicted performance of a *borderline* candidate, were rated *higher* than the PC-value, the observed average item difficulty for all the candidates – *cf.* Table 5.6. This raised further concerns about the validity of the Angoff process as used in this study.

Angoff reliability analysis

Three measures of reliability were used in this study and the Angoff method was found unreliable on all three measures. This was confirmed by similar low IRR ratings across the mini-tests. Only the IRR was used for the tracker questions, since the Angoff process was done five times on the same items across the five test cycles. Similar low/weak IRR's (Kappa's) for the MCQ tracker items were found, which means that the judges were inconsistent their evaluation of the difficulty of the tracker items. These Kappa values suggested that the reliability of the Angoff method was "weak or slight" (Hallgren 2012:5-6). There were too few tracker items to use the other two measures.

5.3.4 MCQ test Cohen data

The 95th percentile points for the performance data of the five cohorts in the full MCQ test ranged from 63 – 77%, with a median value of 71%. This produced Cohen65 pass marks ranging from 41 – 50% (9%), with a median score of 46%. The resulting failure rates were between 32.4 – 51% (18.7%), with a median failure rate of 43%. The two smaller cohorts, January 2013 (n=80) and February 2014 (n=71), both had

Cohen65 pass marks of 41%, whereas the larger three cohorts' pass marks were between 46 - 50% (*cf.* Table 5.4). The 95th percentiles of the two smaller cohorts were similar (63%), but lower than the larger three cohorts (71-77%).

The same trend was also evident in the tracker item data. The Cohen65 pass marks for the five mini-tests ranged from 49 – 55%, with a median score of 54% and resulting failure rates ranged between 43.8 – 65.7% (range= 21.9%), with a median of 51%. The two smaller cohorts, January 2013 (n=80) and February 2014 (n=71), had Cohen65 pass marks of 49-51%, whereas the larger three cohorts' pass marks ranged from 54 – 55% (*cf.* Table 5.5).

This raises the point of whether there were *too few* top performers in the smaller cohorts (Jan 2013 and Feb 2014) with 80 and 71 candidates respectively, thereby influencing (lowering) the Cohen65 pass mark. Although the mean scores were the same across the five cohorts, the smaller SD and DI showed that the top performers in the smaller cohorts did *not* perform to the same high scores as those in the larger cohorts. The validity of the Cohen method, therefore, seems threatened by smaller cohorts and further work is needed to confirm the observation made in this study. A favourable outcome from this validity concern is that the benefit of the doubt about the Cohen method pass mark will go to the candidates if there are too few top performers in a cohort. This is acceptable in the context of the Part I MCQ test since this is not an exit-level test, but one that gives access to higher specialist training.

5.3.5 Comparing the outcomes of the standard setting methods

Review of the *pass marks and failure rates* for the Part I FCP MCQ tests using the two standard setting methods evaluated in this study (from Tables 5.4 and 5.5 as well as Figures 5.5 and 5.6), showed that the range of the Cohen65 pass marks were similar to the Angoff pass marks for the MCQ tests. Although they both had pass mark ranges of 9% across the five test cycles evaluated, they differed by up to 24% on individual cycles. Furthermore, the Angoff failure rates ranged from 65 - 98%, while the Cohen65 failure rates ranged from 32 - 51%.

As seen in Figure 5.5, even the previously used fixed pass mark of 50% performed better than the Angoff method. However, this observation does not justify the use of a fixed pass mark which is not sensitive to variations in test difficulty. Furthermore, the

comparable performance of successive cohorts of test takers included in this study may appear to justify the use of a fixed pass mark. This would be inappropriate because the comparability of test takers and test difficulty is generally not known before a test is administered.

Finally, it is worth noting that the pass mark for the tracker MCQs, which were 30 identical questions, showed little variability using the Cohen method – the Cohen65 pass marks were in a narrow range of 6%. In contrast, the Angoff pass mark varied by up to 26% across the five mini-tests. This measure of inconsistency, on the part of the judges, precludes routine use of the Angoff method, as performed this study.

5.4 CONCLUSION

This chapter presented the results from the second component of the study, specifically relating to the standard setting of the Part I MCQ test. A focussed discussion was also provided, relating to the findings from this chapter. In addition, this chapter contributes to the overall discussion presented in Chapter 8, where conclusions drawn from this component of the study are provided. In the next chapter, Chapter 6, the additional results of the second component of the study, specifically relating to the standard setting of the Part II Objective test, is presented and discussed.

CHAPTER 6

FCP (SA) PART II OBJECTIVE TEST (OT)

COMPARING THE PERFORMANCE OF THE ANGOFF AND COHEN METHODS

6.1 INTRODUCTION

In this chapter, the results of the second component of the study, relating specifically to the Part II OT data, are reported and discussed.

For this part of the study, the results of five consecutive cycles of the written, exit-level FCP (SA) Part II OT were analysed and compared. The results reported here contribute towards answering and addressing the research question (see below) and its related objectives, as discussed in Chapter 3.

Research Question 3: Is using the Cohen method, as compared to Angoff method, a more appropriate way of determining the pass mark for the written assessments of the FCP (SA) examination (Part I MCQ test, Part II Objective test and Part II SEQ test)?

Related objective 1.4.3.4: Determine the performance of the Angoff method of standard setting using five cycles of written FCP (SA) examinations data.

Related objective 1.4.3.5: Determine the performance of the Cohen method of standard setting using the same five cycles of written FCP (SA) examinations data as in 1.4.3.4.

Related objective 1.4.3.6: Determine the variability of the scores of the top-performing candidates, in order to substantiate the assumption within the Cohen method that the test scores of the top candidates sitting the FCP (SA) written examinations, are stable and, therefore, comparable.

Related objective 1.4.3.7: Use the findings of Objectives 1.4.3.2 – 1.4.3.6 to contribute to the evaluation of the utility (as defined in Chapter 2) of the Cohen method, as compared to the Angoff method, for the written FCP (SA) examinations.

Summary of methods

A comprehensive description of the methods used in this study is provided in Chapter 3. However, for convenience, a brief summary is again provided here, mostly to outline the research approach and the data included in the study.

A comparative study design (Altman 1991:6) was used to evaluate the Angoff and Cohen methods of standard setting to answer Research Question 3 and the four related objectives. Five cycles of OT results, a component of the written FCP (SA) examinations, were analysed. Table 3.2 in Chapter 3 describes all the cycles of test data which were included in the study. An abbreviated version of Table 3.2, specific to the test and cycles evaluated in this chapter, is provided to enable easy reference (Table 6.1).

TABLE 6.1: FCP (SA) PART II OT DATA INCLUDED IN THE STUDY

Examination name	Format/Type	Exam cycles
Part II Objective test (OT) (exit exam)	**7-mark OT items x30** constructed-response, short answer question items, focusing on the **interpretation of clinical case scenarios**	Mar 2012
		Aug 2012
		Feb 2013
		Jul 2013
		Feb 2014

The data from the five OT cycles, mentioned in Table 6.1 above, were analysed and are presented in the following order:

- The **candidates' performance data** on the OTs to evaluate the quality and comparability of the papers
- The **Angoff standard setting** outcomes for the OT papers
- The **Cohen standard setting** outcomes for the OT papers

6.2 THE PART II OBJECTIVE TEST (OT) RESULTS

6.2.1 OT candidate performance data

As explained in Chapters 1 and 3, the five cycles of consecutive OTs included in this study all contained 30 constructed-response, 7-mark, short answer items (*cf.* Table 6.1). The total possible score for each paper was 210.

Cohort sizes

The total number of candidates (n) whose data were included in this part of the study was 333. The number of candidates for each test cycle ranged from 57 to 79, with a mean number of candidates per cycle of 67 and a median of 64.

Descriptive statistics (from Table 6.2)

Data in all OT exams were normally distributed according to the Shapiro-Wilk test and Normal Q-Q plots analysis.

The *maximum* scores for the five OT cycles, ranged between 74% - 85% (range= 11%), with a median maximum score of 78%. The *minimum* scores ranged from 25% - 38% (range= 13%), with a median minimum score of 31%.

The *median* of the mean *test scores* across the five OT papers was 56.7% (range= 54.3 – 62.3%). An ANOVA analysis of the mean scores of the five OT cycles showed that there was a statistically significant difference between the OT means ($p<0.001$). Post-hoc Tukey's tests showed that there were statistically significant differences between the mean performance of candidates in *August 2012* (62.3%) vs. March 2012 (56.3%, $p=0.020$); *August 2012* (62.3%) vs. July 2013 (56.7%, $p=0.019$), and *August 2012* (62.3%) vs. February 2014 (54.3%, $p<0.001$).

The *standard deviation* of the means, for the five OT cycles, ranged between 10.1 - 11.0%. The *95% confidence intervals* (CIs) for the mean scores of the cycles are provided in Table 6.2. The width of the CIs across the OT cycles ranged from 4.9% – 5.8%, with a median of 5.2%.

TABLE 6.2: THE PART II OT - PERFORMANCE DATA

Part II Objective Test (OT) exam		Cycles (n= candidates)					Analysis of 5 cycles			
		Mar2012	Aug2012	Feb2013	Jul2013	Feb2014				
Analysis Component	**Analysis Descriptor**	n= 57	n= 64	n= 58	n= 75	n= 79	**Min**	**Max**	**Range**	**Median**
Descriptive statistics	Maximum score (%)	85	83	78	78	74	74	85	11	78
	Minimum score (%)	25	38	30	35	31	25	38	13	31
	Mean score (%)	56.3	62.3	57.0	56.7	54.3	54.3	62.3	8.0	56.7
	Standard Deviation (SD) (%)	11.0	10.4	10.1	10.9	11.0	10.1	11.0	0.9	10.9
	95% confidence intervals (CI) of mean (%)	53.4 - 59.2	59.7 - 64.9	54.4 - 59.6	54.2 - 59.2	51.9 - 56.8	-	-	-	-
	95% CI width (%)	5.8	5.2	5.2	5.0	4.9	4.9	5.8	0.9	5.2
Item analysis	Mean Item Difficulty (ID -value)	0.56	0.62	0.57	0.57	0.54	0.54	0.62	0.08	0.57
	Mean Item Discrimination Index (DI)	0.23	0.23	0.22	0.24	0.25	0.22	0.25	0.03	0.23
	**Item Quality Index (IQI) (%)	69	53	53	71	67	53	71	18	67.0
Test reliability analysis	Cronbach's alpha coefficient	0.89	0.88	0.89	0.89	0.88	0.88	0.89	0.01	0.89
	Standard Error of Measurement (%)	3.6	3.5	3.4	3.6	3.8	3.4	3.8	0.4	3.6

** See text for explanation of IQI

6.2.2 OT item analysis

Item Difficulty value (ID-value)

The overall mean ID-value for all 150 items combined was 0.57. The mean ID-value for all 30 items included in each OT ranged from 0.54 - 0.62, with a median score across the five cycles of 0.57 (*cf.* Table 6.2).

Item Discrimination Index (DI)

The mean DIs of each of the five OT cohorts ranged from 0.22 to 0.25, with a median value of 0.23 (*cf.* Table 6.2). The overall mean DI for all 150 OT items combined was also 0.23.

Item Quality Index (IQI)

The items inside the green zone in Figure of 6.1 were considered good quality items; they had DI's of 0.20 or more and ID-values, expressed as a percentage, between 20% and 80%. Please refer to Appendix E for a detailed explanation of interpreting an Item Quality plot. As shown, the overall IQI for *all* OT items included in the study (n=150), as derived from the Item Quality plot in Figure 6.1, was 63%. The IQI for each OT cycle included in the study ranged from 53% to 71%, with a median value of 67% (*cf.* Table 6.2)

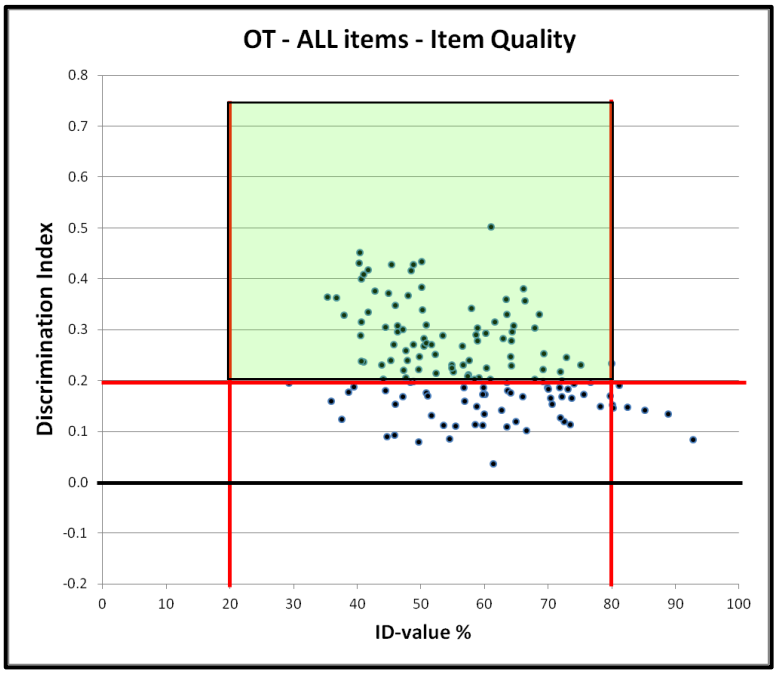


FIGURE 6.1: ITEMS QUALITY PLOT FOR ALL OT ITEMS (n=150)

6.2.3 OT reliability analysis

Cronbach's alpha coefficient for the five OT papers included in the study ranged from 0.88 – 0.89, with a median value of 0.89. The *SEM* of the five test cycles ranged from 3.4 – 3.8, with a median value of 3.6.

6.2.4 OT outcome using the Angoff method

Angoff panel participants

There were 16-18 subject expert judges (CoP examiners) involved in each of the five Part II OT Angoff standard setting procedures. On average, the panellists had 10 years of experience as CoP examiners.

Angoff pass marks

The *Angoff pass marks* generated by the judges for each of the five cycles of the Part II OT ranged from 54 – 66% (range= 12%), with a median of 58% (*cf.* Table 6.3).

Failure rates

The *failure rate* for each OT, based on the pass marks derived using the Angoff method ranged from 35.9% to 79.3% (range= 43.4%), with a median failure rate across the five OT cycles of 49.3% (*cf.* Table 6.3).

In Figure 6.3(a-b), the pass marks and resulting failure rates for the five OT cycles, derived by using the two different standard setting methods evaluated in this study, are shown. The failure rates of the previous fixed 50% pass mark are also included in Figure 6.3b.

TABLE 6.3: THE PART II OT - STANDARD SETTING DATA

Part II Objective Test (OT) exam		Cycles (n= candidates)					Analysis of 5 cycles			
		Mar2012	Aug2012	Feb2013	Jul2013	Feb2014				
Analysis Component	**Analysis Descriptor**	n= 57	n= 64	n= 58	n= 75	n= 79	**Min**	**Max**	**Range**	**Median**
Standard setting analysis and resulting failure rates	95th percentile	73.2	80.0	76.2	73.0	70.2	70	80	10	73.2
	Cohen65 pass mark	**48**	**52**	**49**	**47**	**46**	46	52	6	48.0
	Angoff pass mark (mean of judges)	**54**	**57**	**66**	**58**	**61**	54	66	12	58.0
	Cohen65 FAILURE RATE (%)	21.1	15.6	17.2	17.3	24.1	15.6	24.1	8.4	17.3
	Angoff FAILURE RATE (%)	38.6	35.9	79.3	49.3	67.1	35.9	79.3	43.4	49.3
	50% FAILURE RATE (%)	24.6	9.4	22.4	24.0	35.4	9.4	35.4	26.1	24.0
Angoff validity analysis	Pearson Correlation - Angoff % vs ID-value r=	*0.44*	0.29	*0.55*	*0.47*	*0.60*	0.29	0.60	0.31	0.47
	Correlation - Angoff % vs ID-value r^2=	0.19	0.08	0.31	0.22	0.36	0.08	0.36	0.27	0.22
Angoff method reliability analyses	SEMean Angoff pass mark	1.6	2.1	3.1	1.7	2.0	1.6	3.1	1.5	2.0
	Max SEMean allowed for reliable Angoff result[1]	1.8	1.8	1.7	1.8	1.9	1.7	1.9	0.2	1.8
	Standard deviation (SD) of Angoff pass marks	6.5	8.4	12.2	7.3	7.8	6.5	12.2	5.7	7.8
	Max Angoff SD allowed for reliable result[2]	2.7	2.6	2.5	2.7	2.7	2.5	2.7	0.2	2.7
	Inter-rater Reliability - Intra-class Correlations[3]	0.75	0.71	0.81	0.66	0.76	0.66	0.81	0.15	0.75
	Intra-class Correlations - 95% confidence intervals	0.60 - 0.86	0.55 - 0.84	0.59 - 0.95	0.45 - 0.83	0.62 - 0.87	-	-	-	-
	ICC 95% confidence intervals width	0.26	0.29	0.36	0.38	0.25	0.25	0.38	0.13	0.29

SEMean = Standard Error of the Mean *Red italic numbers: Correlations were statistically significant (p<0.020)*

[1] Cohen, Kane, Crooks 1999

[2] Meskauskas 1986

[3] Hallgren 2012

Validity of the Angoff method

A *validity-correlation plot* (*cf.* Chapter 2), was used as a quantitative measure of the internal validity of the Angoff procedure. This was done by determining the Pearson correlation coefficient (r), for the relationship between the predicted test item difficulty (Angoff rating expressed as a percentage) and actual test item difficulty (ID-value expressed as a percentage) of the individual OT items. Figure 6.2 is the validity-correlation plot for all 150 OT items used over the five cycles. The Pearson correlation coefficient (r) was 0.37, which was statistically significant (p<0.001). The corresponding r^2 value was 0.14.

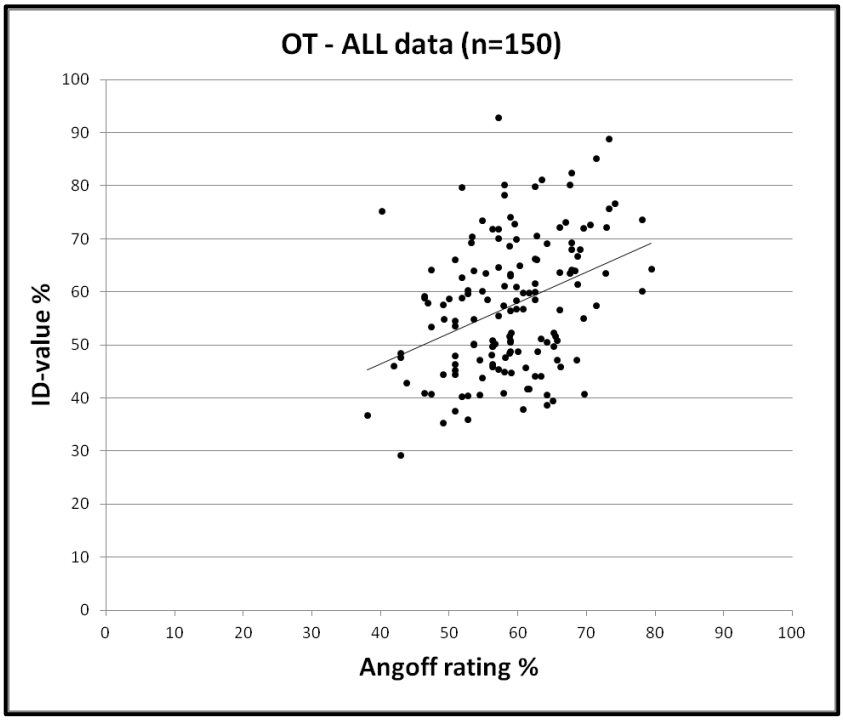


FIGURE 6.2: VALIDITY CORRELATION PLOT FOR ALL OT ITEMS (n=150)

The Pearson correlation values (r) for each individual OT ranged from 0.29 to 0.60, with a median of 0.47 (*cf.* Tables 6.3 or 6.4) and corresponding r^2 values from 0.08 to 0.36 (*cf.* Table 6.3). All the correlation coefficients were statistically significant (p <0.020), except the 0.29 value (p =0.12).

Table 6.4 below summarises the Angoff data from each OT cycle in terms of three indicators:

- Angoff method pass mark (%);
- Mean ID-value (%);
- Pearson correlation coefficient (r) between the two parameters listed above.

Table 6.4 shows that the Angoff pass mark percentage, which is the predicted performance of a *borderline* candidate, was *higher* in the last three test cycles (February 2013, July 2013 and February 2014) than the mean ID-value percentage, which is a reflection of the performance of the *average* candidate. In the other two cycles, the mean ID-value, expressed as a percentage, was higher than the Angoff pass mark percentage, which is the expected outcome – that an average candidate will score higher than a borderline candidate.

TABLE 6.4: THE PART II OT: ANGOFF ANALYSIS SUMMARY

Cycle (candidates)	**Angoff pass mark** (%)	**Mean ID-value** (%)	**Pearson Correlation** (r)	**r^2**
Mar 2012 (57)	54	56.4	*0.44*	0.19
Aug 2012 (64)	57	62.3	0.29	0.08
Feb 2013 (58)	66	57.0	*0.55*	0.31
Jul 2013 (75)	58	56.6	*0.47*	0.22
Feb 2014 (79)	61	54.3	*0.60*	0.36
Median	**58**	**57.0**	**0.47**	**0.22**
Range	**12**	**8**	**0.31**	**0.27**

Red italic numbers: Correlations were statistically significant, p-values <0.020

Reliability of the Angoff method

Three indicators of reliability were used and triangulated in this study to form a judgement regarding the reliability of the Angoff process used for the OT cycles. The data from these three measures are provided in Table 6.3.

The first method used, described by Cohen, Kane, Crooks (1999), showed two of the five cycles (March 2012 and July 2013) had a reliable outcome according to this method (highlighted green), the other three cohorts did not have a reliable result and was highlighted red – *cf.* Table 6.3.

The second method used, described by Meskauskas (1986), showed that none of the five cycles had a reliable outcome and hence was highlighted red – *cf.* Table 6.3.

The third and probably the most robust reliability estimation method which was used in this study, was the *inter-rater reliability* (IRR) calculation. The specific form of IRR that was used was the Intra-class Correlation (ICC) coefficient method due to the scale data of the Angoff ratings and since there were more than two raters [cf. Hallgren (2012:5-6)]. The ICC coefficients across the five cycles were between 0.66 – 0.81, with 95% confidence intervals (CIs) of between 0.25 – 0.38 per cycle (*cf.* Table 6.3). The interpretation of the IRR outcome is addressed in the discussion section.

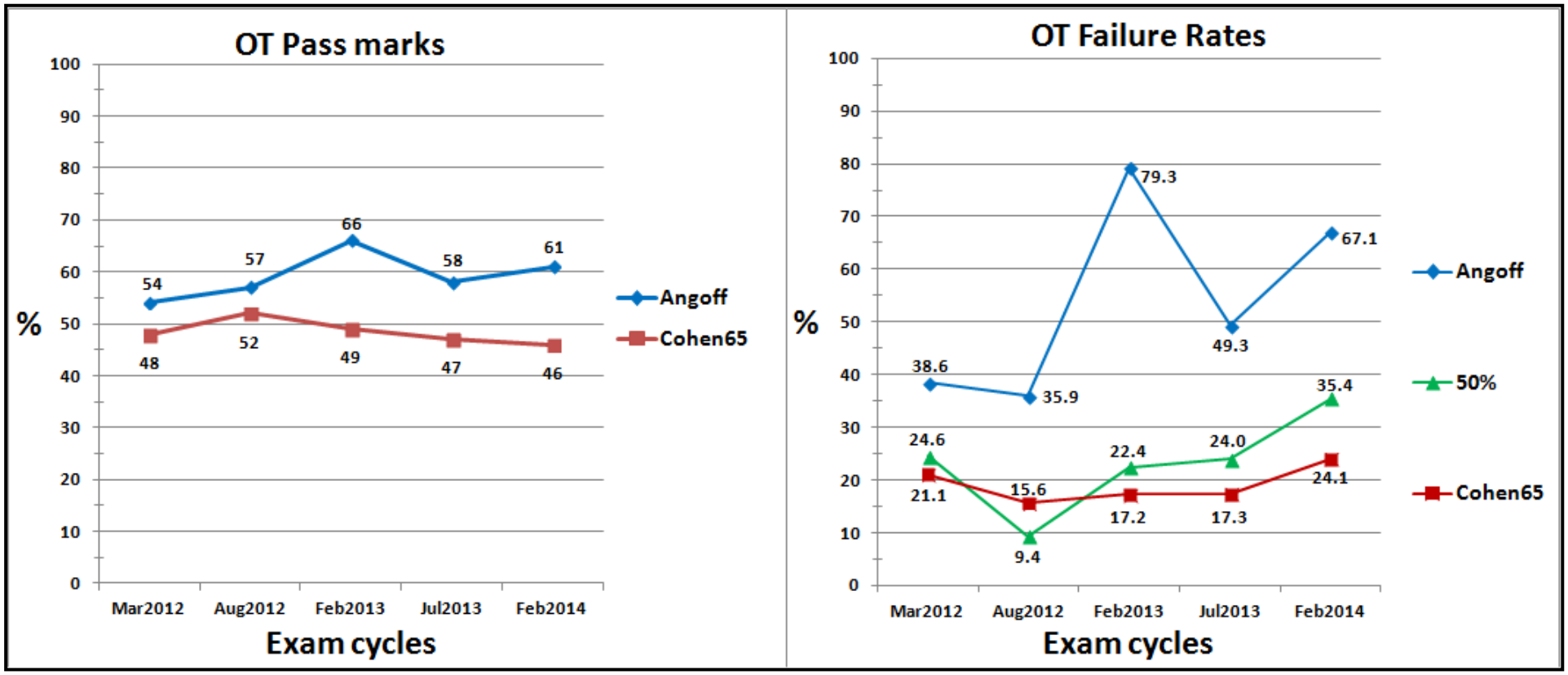


FIGURE 6.3(a-b): OT PASS MARKS (a) AND RESULTING FAILURE RATES (b)

6.2.5 OT outcome using the Cohen method

The results of the Cohen method of standard setting (Cohen65 model) for the five OT cycles, included in this study, are provided in Table 6.3. The *Cohen65 pass marks* for the five cohorts ranged from 46 – 52% (6%), with a median of 48%.

The resulting *failure rates* ranged between 15.6 – 24.1% (8.4%), with a median of 17.3% (*cf.* Table 6.3).

The *95th percentiles*, from which the Cohen65 pass marks were derived, ranged from 70 – 80%, with a median of 73.2% (*cf.* Table 6.3).

The pass marks and resulting failure rates of the Angoff and Cohen methods are graphically expressed for the OT papers in Figure 6.3(a-b). The failure rates of the previous fixed 50% pass mark are also included in Figure 6.3b.

6.2.6 OT outcome using a 50% pass mark

The previous traditional *fixed 50% pass mark* practice would have resulted in failure rates of between 9.4 – 35.4% (range= 26.1%), with a median failure rate of 24% for the five OT cycles included in the study (*cf.* Table 6.3).

6.3 THE PART II OBJECTIVE TEST DISCUSSION

This chapter contributes data towards addressing Research Question 3: "Is using the Cohen method, as compared to Angoff method, a more appropriate way of determining the pass mark for the written assessments of the FCP (SA) examination (Part I MCQ test, Part II Objective test and Part II SEQ test)?"

In this chapter, results of the second component of the study, relating specifically to the entry-level FCP (SA) Part II Objective Test (OT) performance data, were analysed and reported. Data from five consecutive OT cycles were used. The performance of the Angoff and Cohen methods of standard setting, using the data described, were analysed, reported and will now be compared during the discussion of their respective outcomes.

The discussion is structured broadly in the same order as the results presentation in the preceding text.

6.3.1 Candidates and Cohorts

The criteria to gain entry to sit the Part II exit examination were explained in Chapter 1 and 3. For these selection reasons the exit-level cohorts are more homogeneous, compared to the Part I MCQ tests' cohorts. Similar to the Part I cohorts, there were no additional candidate information available to understand the formation of the different cohorts better. The need for more comprehensive candidate information is well recognised.

The number of Part II exit examination candidates is steadily increasing (*cf.* Table 6.2). This is due to the single exit examination system, introduced by the HPCSA in 2011, starting to take effect (*cf.* Chapter 1). Compared to data from 2001 – 2005, when there were on average 24 candidates per FCP (SA) Part II sitting (Burch-papers – *cf.* Chapter 2), in this present study the mean number of candidates was 67 and the median 64. This trend will probably continue due to the 2010 HPCSA single exit-examination rule taking effect, as discussed in Chapter 1.

6.3.2 OT candidate performance data

The maximum and minimum ranges in the OT cohorts were similar, at 11% and 13% respectively. This was probably due to the greater homogeneity of the cohorts and that there is no negative marking.

An ANOVA analysis of the mean scores of the five OT cycles showed that there was a statistically significant difference between the OT means ($p<0.001$). Post-hoc tests showed that there were statistically significant differences between the mean performances of candidates in August 2012 (62.3%) vs. March 2012 (56.3%, $p=0.020$); August 2012 (62.3%) vs. July 2013 (56.7%, $p=0.019$); August 2012 (62.3%) vs. February 2014 (54.3%, $p<0.001$). This finding suggests that the August 2012 paper was easier than the other four OT cycles.

The standard deviation (SD) across all the cohorts were in a very narrow range of 0.9%, which indicates that the cohorts had similar distributions around their means,

including the August 2012 paper, whose mean was different from the three other papers as mentioned above. The 95% Confidence intervals (CIs) were also narrow, so there can be a high level of confidence in the mean scores.

OT item analysis

The mean ID-value per cycle, converted to % is the same value as the performance means per cohort. Therefore, the ID-value discussion is similar to that of the mean performance score.

The mean DIs of the 30 items of each cohort were above 0.20 and therefore of acceptable quality from an item discrimination ability perspective. They ranged from 0.22 to 0.25, with a median of 0.23.

The composite IQI percentage for the overall 150 OT items was 63% and across the five cycles the IQI ranged from 53 - 71%, which reflected a good and acceptable quality written test.

Reliability analysis

The OT papers had excellent and stable reliability measures across all five cycles. Cronbach's alphas ranged from 0.88 – 0.89 with a median of 0.89. The SEMs of the five cycles ranged from 3.4 – 3.8 with a median of 3.6.

6.3.3 OT Angoff data

There were 16-18 subject expert judges involved in each of the five Part II OT Angoff standard setting procedures, which is adequate according to the presented literature. The mean duration of involvement in the FCP (SA) examinations of the panellists was 10.0 years. This indicated that experienced clinician judges were involved in the Angoff processes. The content and candidates of the OTs are well known to the judges, since it is an exit-level exam based on clinical cases.

The Angoff pass marks generated by the judges for the five cycles of the Part II OT ranged from 54 – 66% (12%), with a median of 58% (*cf.* Table 6.3). The resulting failure rates which these Angoff pass marks produced ranged from 35.9% to 79.3% (43.4%), with a median failure rate across the five OT cycles of 49.3% (*cf.* Table 6.3). The previous traditional fixed 50% pass mark practice would have resulted in failure

rates of between 9.4 – 35.4% (26.1%), with a median failure rate of 24% for the five OT cycles included in the study (*cf.* Table 6.3).

The failure rates were very high for an exit-exam across the five OT cycles (36-79%) and had a wide range of 43%, although candidates' performances over the five cycles were broadly similar in terms of means and SDs, except for the August 2012 cycle, which appeared easier. Despite the similar performances of the cohorts in the OT, this was not reflected in the Angoff pass marks. The easiest OT (August 2012) had the second lowest Angoff pass mark and was rated by the judges as the second hardest, which was concerning.

Angoff validity analysis

The Pearson correlation (r) between the predicted (Angoff rating %) and actual difficulty (ID-value %) across all the OT 150 items was 0.37 (p <0.001) and the corresponding r^2 value 0.14. This was a weak correlation and only 14% of the relationship can be explained by the data, the rest is random error in the judgement process. This was unexpected since examiners probably had better understanding of the abilities of the exit-level candidate than of the entry-level and the OT test is more clinically focussed than the basic science based MCQ. Both the model answers were provided during the Angoff process, but the judges struggled to identify the hard from the easy items for the borderline competent student.

However, the correlations within the five individual OT papers were stronger and ranged from 0.29 (weak) to 0.60 (moderate) and the corresponding r^2 values from 0.08 to 0.36 (*cf.* Table 6.3). All the correlations were statistically significant (p <0.020), except the 0.29 value (p =0.12). This meant that 8 - 36 % of the relationship between the predicted (Angoff rating %) and actual difficulty (ID-value %) within respective OT papers can be explained by the data, the rest is random error in the judgement process. This was a significant validity concern.

Angoff reliability analysis

Three measures of reliability were again used in this study. The Angoff method had weak reliability over all five cohorts according to the Meskauskas method of evaluating the SD of the Angoff pass marks. The Cohen, Kane and Crooks method showed that the March 2012 and July 2013 judgements were reliable, but the other three cohorts did not meet the standard. The final method was calculating the IRR, using the ICC,

which showed that the ICC point values indicated good to excellent (highlighted green) reliability of the Angoff process of the judges (Hallgren 2012:5-6), but the wide CIs made these ICC values less useful.

Therefore, reliability was a concern in four out of five cohorts and only the March 2012 had acceptable reliability. The wide CIs of the ICC IRR values rendered them much less useful, especially in the context of how Angoff's reliability was reflected by the other two methods, where only two out of ten measurements came back as reliable.

6.3.4 OT Cohen data

The 95th percentile range of 10% for the Cohen method across the five cycles of OT data, and the resultant Cohen65 range of 6% can be described as a stable performance of the Cohen method, if considering the fact that the August 2012 cycle was a statistically significantly easier test, compared to the other four cycles. The range of Cohen65 pass marks was half the range of the Angoff pass marks (12%). Excluding the easier cycle (August 2012) results in a Cohen65 pass mark range of only 3%, whereas the Angoff pass mark range remains unchanged at 12% (*cf.* Table 6.3). This reflects the lack of the Angoff method, as used in this study, to respond to test difficulty. The pass marks of the Angoff and Cohen65 methods did not overlap on any paper (*cf.* Figure 6.3)

The validity of the Cohen method did not appear threatened by smaller OT cohorts, as was the case in the MCQ test cohorts. Cohort size and SD had no apparent effect on the 95th percentiles like in the MCQ test, although the SDs were virtually the same across the five OT cohorts. The most likely reasons for this finding are the homogeneity of exit-level candidates, in terms of examination preparedness, and the consistency (reliability) of the papers.

6.3.5 Comparing the outcomes of the standard setting methods

On reviewing the pass marks and failure rates of the two methods used in this study (from Tables 6.3 and Figures 6.3):

The Angoff pass marks were consistently higher than the Cohen65 pass marks on *all* OT cycles. Their respective pass marks differed by between 5 - 17% on individual

cycles. The respective failure rates which they produced showed considerable differences. Angoff failure rates had major variation and ranged from approximately 36 - 79% and Cohen65 failure rates ranged from approximately 15 - 24%. If Angoff was operational in the CoP it would have led to mass failures and most likely mass appeals from candidates and subsequent CMSA intervention. The Cohen65 failure rates were much more acceptable to the CoP and were accepted and ratified by the CMSA.

To put the unrealistic and unacceptable Angoff failure rates in context, the previous fixed 50% pass mark practice, which was indefensible per se, yielded consistently lower and more stable failure rates than the Angoff method, which were probably still within the acceptable spectrum of failure rates. The Angoff failure rates did not overlap with any of the 50% or Cohen65 failure rates in any cycle and was clearly unrealistic. The 50% failure rates tracked the Cohen65 failure rates and overlapped twice over the five cohorts.

6.4 CONCLUSION

This chapter presented the results from the second component of the study, specifically relating to the standard setting of the Part II Objective test. A focussed discussion was also provided, relating to the findings from this chapter. In addition, this chapter contributes to the overall discussion presented in Chapter 8, where conclusions drawn from this component of the study are provided. In the next chapter, Chapter 7, the additional results of the second component of the study, specifically relating to the standard setting of the Part II SEQ test, is presented and discussed.

CHAPTER 7

FCP (SA) PART II SHORT ESSAY QUESTION (SEQ) TEST

COMPARING THE PERFORMANCE OF THE ANGOFF AND COHEN METHODS

7.1 INTRODUCTION

In this chapter, the results of the second component of the study, relating specifically to the Part II SEQ test data, are reported and discussed.

For this part of the study, the results of five consecutive cycles of the written, exit-level FCP (SA) Part II SEQ test were analysed and compared. The results reported here contribute towards answering and addressing the research question (see below) and its related objectives, as discussed in Chapter 3.

Research Question 3: Is using the Cohen method, as compared to Angoff method, a more appropriate way of determining the pass mark for the written assessments of the FCP (SA) examination (Part I MCQ test, Part II Objective test and Part II SEQ test)?

Related objective 1.4.3.4: Determine the performance of the Angoff method of standard setting using five cycles of written FCP (SA) examinations data.

Related objective 1.4.3.5: Determine the performance of the Cohen method of standard setting using the same five cycles of written FCP (SA) examinations data as in 1.4.3.4.

Related objective 1.4.3.6: Determine the variability of the scores of the top-performing candidates, in order to substantiate the assumption within the Cohen method that the test scores of the top candidates sitting the FCP (SA) written examinations, are stable and, therefore, comparable.

Related objective 1.4.3.7: Use the findings of Objectives 1.4.3.2 – 1.4.3.6 to contribute to the evaluation of the utility (as defined in Chapter 2) of the Cohen method, as compared to the Angoff method, for the written FCP (SA) examinations.

Summary of methods

A comprehensive description of the methods used in this study is provided in Chapter 3. However, for convenience, a brief summary is again provided here, mostly to outline the research approach and the data included in the study.

A comparative study design (Altman 1991:6) was used to evaluate the Angoff and Cohen methods of standard setting to answer Research Question 3 and the four related objectives. Five cycles of SEQ test results, a component of the written FCP (SA) examinations, were analysed. Table 3.2 in Chapter 3 describes all the cycles of test data which were included in the study. An abbreviated version of Table 3.2, specific to the test and cycles evaluated in this chapter, is provided to enable easy reference (Table 7.1).

TABLE 7.1: FCP (SA) PART II SEQ TEST DATA INCLUDED IN THE STUDY

Examination name	Format/Type	Exam cycles
Part II Short Essay Question (SEQ) test (exit exam)	**15-mark SEQ items x20** constructed-response **theory** short essay questions	Aug 2011
		Mar 2012
		Aug 2012
		Feb 2013
		Jul 2013

The SEQ tests' data were analysed and are presented in the following order:

- The **candidates' performance data** on the SEQ tests, to evaluate the quality and comparability of the papers
- The **Angoff standard setting** outcomes for the SEQ test papers
- The **Cohen standard setting** outcomes for the SEQ test papers

7.2 THE PART II SHORT ESSAY QUESTION (SEQ) TEST RESULTS

Variability of SEQ test data

As explained in Chapters 1 and 3, five cycles of consecutive SEQ tests were used. Each test consisted of 20 constructed-response, 15-mark short answer items (*cf.* Table 7.1). The total possible score for the SEQ test was 300.

The SEQ test data evaluated in this component of the study were variable in format. The raw score for each question in each test was made up of two sub-questions, each scored out of 15 marks. Many markers combined the scores of the two 15-mark items and only submitted a single score out of 30. To enable uniform analysis of the SEQ test data, the score for each question, consisting of two 15-mark sub-questions, was combined to form 30-mark test items. This meant that each SEQ test was analysed as a test consisting of ten 30-mark items, with a maximum possible score of 300 marks.

7.2.1 SEQ candidate performance data

Cohort sizes

The total number of candidates sitting the five SEQ tests included in the study was 309. The number of candidates sitting each test ranged from 55 – 75 (*cf.* Table 7.3), with a mean number of candidates per cycle of 62 and a median of 58.

Descriptive statistics (from Table 7.3)

Data in all the SEQ tests were normally distributed according to the Shapiro-Wilk test and Normal Q-Q plots analysis.

The *maximum* scores for the five SEQ tests evaluated, ranged from 62% to 75%, with a median score of 72%. The *minimum* scores ranged from 29% to 43%, with a median score of 38%.

The five *mean performance* scores of the individual SEQ papers had a median of 56.1% (range= 46.3 – 57.9%). An ANOVA analysis showed that the mean scores of the respective SEQ tests were significantly different ($p<0.001$). Post-hoc Tukey's tests showed that the *February 2013* SEQ test had a statistically significantly lower mean score than all the other SEQ tests included in this study (*cf.* Table 7.2).

TABLE 7.2: SIGNIFICANT DIFFERENCES IN THE MEAN SEQ PERFORMANCES

SEQ cohort pairs (mean %)		***p* - value**
Mar 2012 (57.9)	Aug 2012 (53.1)	$p = 0.001$
Feb 2013 (46.3)	Aug 2011 (56.4)	$p \leq 0.001$
Feb 2013 (46.3)	Mar 2012 (57.9)	$p \leq 0.001$
Feb 2013 (46.3)	Aug 2012 (53.1)	$p \leq 0.001$
Feb 2013 (46.3)	Jul 2013 (56.1)	$p \leq 0.001$

The *standard deviation* around the means, across the five cycles, had a narrow range between 6.5 - 7.3% (median = 7.0%). The *95% confidence intervals* (CIs) for the individual means of the five tests are provided in Table 7.3. The width of the CIs across the SEQ test cycles ranged from 3.1% – 3.8%, with a median of 3.6%.

TABLE 7.3: THE PART II SEQ TEST - PERFORMANCE DATA

Part II Short Essay Question (SEQ) exam		Cycles (n= candidates)					Analysis of 5 cycles			
		Aug2011	Mar2012	Aug2012	Feb2013	Jul2013				
Analysis Component	Analysis Descriptor	n= 55	n= 57	n= 64	n= 58	n= 75	Min	Max	Range	Median
Descriptive statistics	Maximum score (%)	70	74	72	62	75	62	75	13	72
	Minimum score (%)	35	43	38	29	41	29	43	14	38
	Mean score (%)	56.4	57.9	53.1	46.3	56.1	46.3	57.9	11.7	56.1
	Standard Deviation (SD) (%)	7.0	6.5	7.3	7.1	6.7	6.5	7.3	0.7	7.0
	95% confidence intervals (CI) of mean (%)	54.5 - 58.3	56.2 - 59.7	51.3 - 54.9	44.4 - 48.1	54.5 - 57.6	-	-	-	-
	95% CI width (%)	3.8	3.5	3.6	3.7	3.1	3.1	3.8	0.7	3.6
Item analysis	Mean Item Difficulty (ID -value)	0.56	0.58	0.53	0.46	0.56	0.46	0.58	0.12	0.56
	Mean Item Discrimination Index (DI)	0.15	0.14	0.15	0.16	0.14	0.14	0.16	0.02	0.15
	**Item Quality Index (IQI) (%)	20	10	20	20	0	0	20	20	20
Test reliability analysis	Cronbach's alpha coefficient	0.78	0.76	0.77	0.78	0.79	0.76	0.79	0.03	0.78
	Standard Error of Measurement (%)	3.3	3.2	3.5	3.3	3.0	3.0	3.5	0.5	3.3

** See text for explanation of IQI

7.2.2 SEQ test item analysis

Item Difficulty value (ID-value)

The overall mean ID-value for all 50 SEQ test items included in the five SEQ tests, was 0.54. The mean ID-values for each SEQ test's ten test items were between 0.46 - 0.58, with a median score across the five cycles of 0.56 (*cf.* Table 7.3).

Item Discrimination Index (DI)

The mean DI of each of the five SEQ tests ranged from 0.14 to 0.16, with a median of 0.15 (*cf.* Table 7.3).

Item Quality Index (IQI)

The items inside the green zone of Figure 7.1 were considered good quality items. They had DI's of 0.20 or more and ID-value % values between 20% and 80%. Please refer to Appendix E for a detailed explanation of the interpretation of an Item Quality plot. The overall IQI for *all* the SEQ test items included in the study (n=50), as derived from the Item Quality plot in Figure 7.1, was 14%. The IQI for each SEQ test in the study ranged from 0 - 20%, with a median 20% (*cf.* Table 7.3)

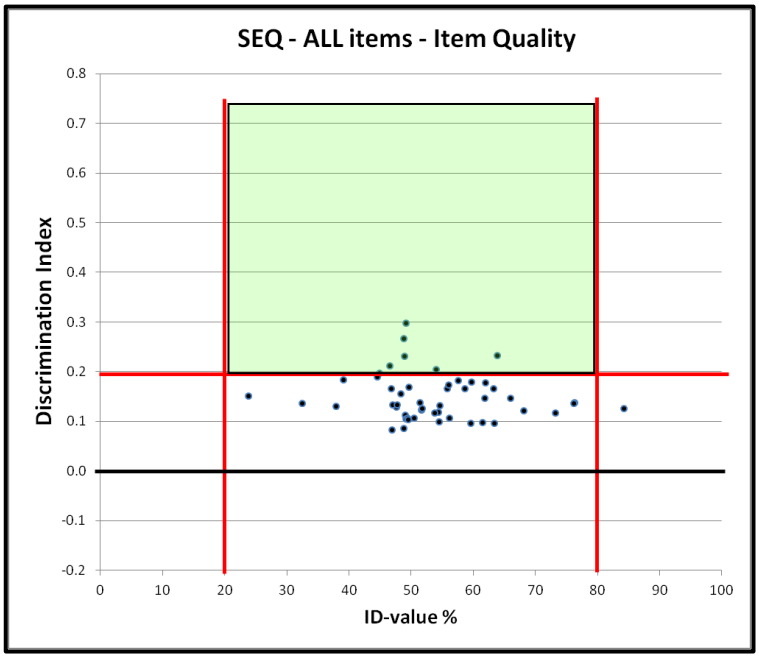


FIGURE 7.1: ITEM QUALITY PLOT FOR ALL SEQ ITEMS (n=50)

7.2.3 SEQ test reliability analysis

Cronbach's alpha coefficients for the five SEQ tests included in the study ranged between 0.76 – 0.79, with a median value of 0.78. The *SEM* of the five SEQ test cycles ranged from 3.0 – 3.5, with a median of 3.3.

7.2.4 SEQ test outcome using the Angoff method

Angoff panel participants

There were 11 - 16 subject expert judges (CoP examiners) involved in the Angoff standard setting procedure for each of the five SEQ tests included in this study. On average, the panellists had been FCP examiners for 11.7 years.

Pass marks

The *Angoff pass marks* generated by the judges for the five cycles of the SEQ tests ranged from 50 – 52%, with a median of 52% (*cf.* Table 7.4).

Failure rates

The failure rates for each SEQ test, using the pass mark derived from the Angoff procedure ranged from 17.5% to 70.7% (range= 53.1%), with a median failure rate across the five SEQ test cycles of 26.7% (cf. Table 7.4).

Figure 7.3(a-b) shows the pass marks and resulting failure rates for the five SEQ test cycles, using the two different standard setting methods evaluated in this study. The failure rates of the previous fixed 50% pass mark are also included in Figure 7.3b.

TABLE 7.4: THE PART II SEQ TESTS - STANDARD SETTING DATA

Part II Short Essay Question (SEQ) exam		Cycles (n= candidates)					Analysis of 5 cycles			
		Aug2011	Mar2012	Aug2012	Feb2013	Jul2013				
Analysis Component	**Analysis Descriptor**	n= 55	n= 57	n= 64	n= 58	n= 75	**Min**	**Max**	**Range**	**Median**
Standard setting analysis and resulting failure rates	95th percentile	68.3	69.2	66.9	58.2	67.0	58	69	11	67.0
	Cohen65 pass mark	**44**	**45**	**43**	**38**	**44**	38	45	7	44.0
	Angoff pass mark (mean of judges)	**52**	**52**	**50**	**50**	**52**	50	52	2	52.0
	Cohen65 FAILURE RATE %	3.6	1.8	4.7	8.6	4.0	1.8	8.6	6.9	4.0
	Angoff FAILURE RATE %	20.0	17.5	34.4	70.7	26.7	17.5	70.7	53.1	26.7
	50% FAILURE RATE %	18.2	7.0	34.4	70.7	16.0	7.0	70.7	63.7	18.2
Angoff validity analysis	Pearson Correlation - Angoff % vs ID-value r=	-0.07	-0.35	0.17	*0.78*	-0.10	-0.35	0.78	1.12	-0.07
	Correlation - Angoff % vs ID-value r^2=	0.01	0.12	0.03	0.60	0.01	0.01	0.60	0.60	0.03
Angoff method reliability analyses	SEMean Angoff pass mark	1.6	1.9	1.9	2.7	2.1	1.6	2.7	1.1	1.9
	Max SEMean allowed for reliable Angoff result[1]	1.7	1.6	1.8	1.7	1.5	1.5	1.8	0.3	1.7
	Standard deviation (SD) of Angoff pass marks	5.4	6.5	6.7	8.9	8.3	5.4	8.9	3.5	6.7
	Max Angoff SD allowed for reliable result[2]	1.7	1.6	1.8	1.8	1.7	1.6	1.8	0.2	1.7
	Inter-rater Reliability - Intra-class Correlations[3]	0.90	0.72	0.68	0.51	0.64	0.51	0.90	0.39	0.68
	Intra-class Correlations - 95% confidence intervals	0.81 - 0.95	0.51 - 0.87	0.45 - 0.85	0.23 - 0.74	0.41 - 0.82	-	-	-	-
	ICC 95% confidence intervals width	0.14	0.36	0.4	0.51	0.41	0.14	0.51	0.37	0.40

SEMean = Standard Error of the Mean *Red italic numbers: Correlations were statistically significant (p=0.008)*

[1] Cohen, Kane, Crooks 1999

[2] Meskauskas 1986

[3] Hallgren 2012

Validity of the Angoff method

A *validity-correlation plot* (*cf.* Chapter 2), was used as a quantitative measure of the internal validity of the Angoff procedure. This was done by determining the Pearson correlation coefficient (r), for the relationship between the predicted test item difficulty (Angoff rating expressed as percentage) and actual test item difficulty (ID-value expressed as a percentage) of the individual SEQ test items. Figure 7.2 is the validity-correlation plot for all 50 SEQ test items used over the five test cycles. The Pearson correlation coefficient (r) was 0.10, which was not statistically significant (p =0.483) and the corresponding r^2 value was 0.01.

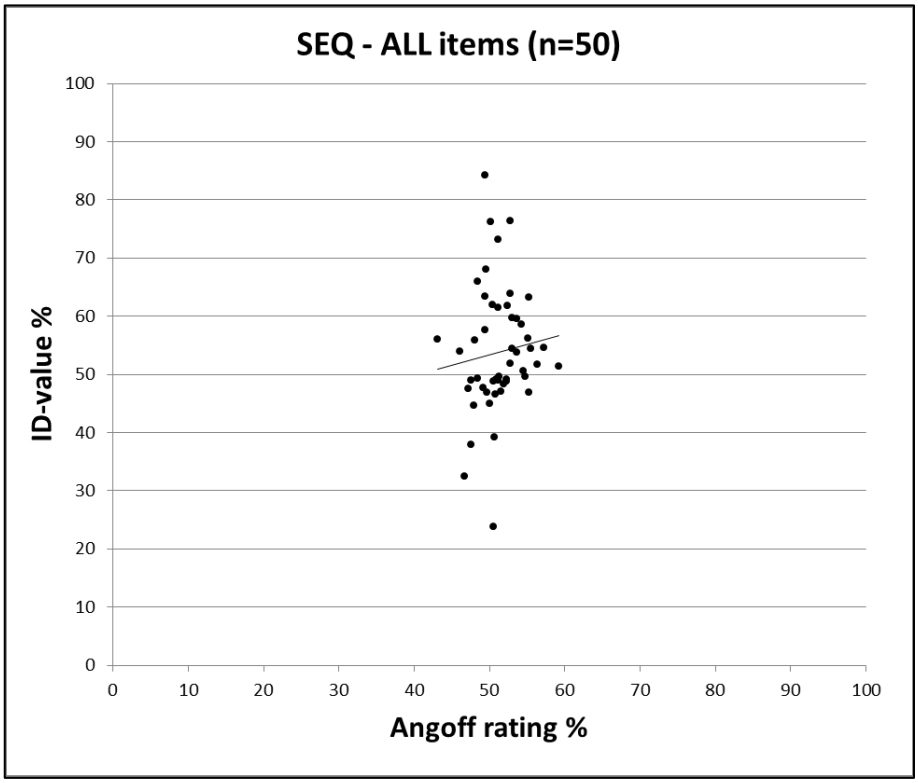


FIGURE 7.2: VALIDITY CORRELATION PLOT FOR ALL SEQ TEST ITEMS (n=50)

The Pearson correlation (r) for the ten items in each of the five individual SEQ test papers ranged from -0.35 to 0.78, with a median of -0.07. The corresponding r^2 values ranged from 0.01 to 0.60 (*cf.* Table 7.4 and 7.5). Only *one* of the correlations (February 2013, r= 0.78) was statistically significant (p= 0.008).

Table 7.5 below summarises the Angoff data from each SEQ test cycle in terms of three indicators:

- Angoff method pass mark (%);
- Mean ID-value (%);
- Pearson correlation coefficient (r) between the two parameters listed above.

Table 7.5 shows that the Angoff pass mark, i.e. the predicted performance of a *borderline* candidate, was *higher* in one test cycle (February 2013) than the mean ID-value %, i.e. the real performance of the *average* candidate. In the other four cycles, the mean ID-value % was higher than the Angoff pass mark %, which is expected – that the average candidate will score higher than the borderline candidate.

TABLE 7.5: THE PART II SEQ TESTS: ANGOFF ANALYSIS SUMMARY

Cycle (n)	**Angoff pass mark** (%)	**Mean ID-value** (%)	**Pearson Correlation** (r)	r^2
Aug 2011 (55)	52	56.4	-0.07	0.01
Mar 2012 (57)	52	57.8	-0.35	0.12
Aug 2012 (64)	50	53.1	0.17	0.03
Feb 2013 (58)	50	46.3	*0.78*	0.60
Jul 2013 (75)	52	56.1	-0.10	0.01
Median	**52**	**56.1**	**-0.07**	**0.03**
Range	**2**	**11.7**	**1.12**	**0.60**

Red italic number: Correlation was statistically significant, p-value = 0.008

Reliability of the Angoff method

Three indicators of reliability were used and triangulated in this study to form a judgement regarding the reliability of the Angoff process used for the SEQ tests. These three methods are described in detail in Chapter 3. The data from these three measures are provided in Table 7.4.

The first method used, described by Cohen et al. (1999), shows that *one* of the five cycles (August 2011) had a reliable outcome according to this method (highlighted in green), while the other four tests did not (highlighted in red - *cf.* Table 7.4).

The second method used, described by Meskauskas (1986), shows that *none* of the five cycles had a reliable outcome (highlighted in red - *cf.* Table 7.4).

The third, and probably the most robust reliability estimation method used in this study, was the *inter-rater reliability* (IRR). The specific measure of IRR used in this study was the Intra-class Correlation (ICC) coefficient, due to the scale (or interval) data type of the Angoff ratings (0-30 marks) and because more than two raters scored each item [*cf.* Hallgren (2012:5-6)]. The ICC coefficients across the five cycles ranged from 0.51 – 0.90, with wide 95% confidence intervals (CIs) of between 0.14 – 0.51 per cycle (*cf.* Table 7.4). The interpretation of the IRR is addressed in the discussion section.

7.2.5 SEQ test outcome using the Cohen method

The results of the Cohen method of standard setting (Cohen65 model) for the five SEQ test cycles, included in this study, are provided in Table 7.4. The *Cohen65 pass marks* for the five cohorts ranged from 38 – 45%, with a median of 44%.

The resulting *failure rates* ranged between 1.8 – 8.6% (range= 6.9%), with a median of 4.0% (*cf.* Table 7.4).

The *95th percentiles*, from which the Cohen65 pass marks were derived, ranged from 58.2 – 69.2%, with a median of 67.0% (*cf.* Table 7.4).

The pass marks and corresponding failure rates derived using the Angoff and Cohen methods are graphically expressed for the SEQ tests in Figure 7.3(a-b).

7.2.6 SEQ test outcome using a 50% pass mark

The previous traditional fixed 50% pass mark practice would have resulted in failure rates between 7.0 – 70.7% (range= 63.7%), with a median failure rate of 18.2% for the five SEQ test cycles included in the study (cf. Table 7.4).

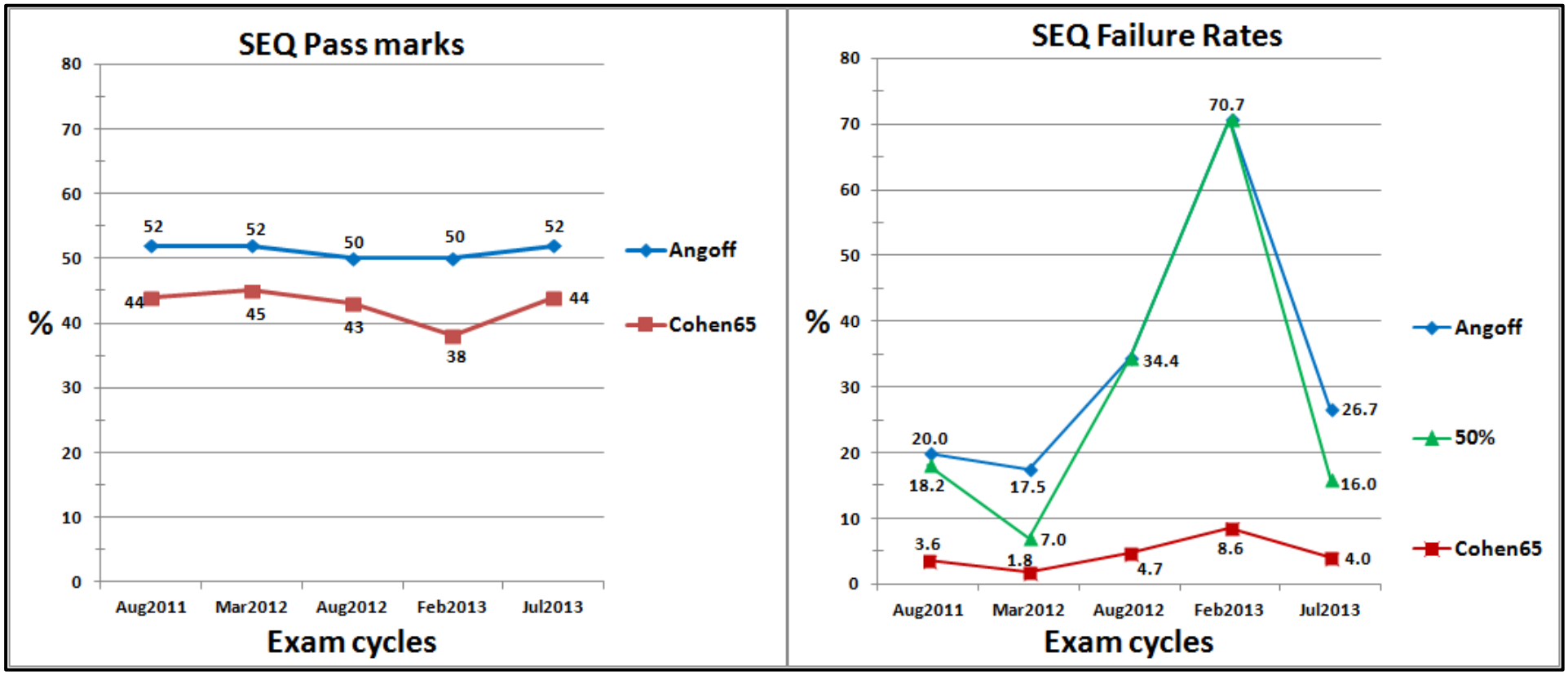


FIGURE 7.3(a-b): SEQ PASS MARKS (a) AND RESULTING FAILURE RATES (b)

7.3 THE PART II SEQ TEST DISCUSSION

This chapter contributes data towards addressing Research Question 3: "Is using the Cohen method, as compared to Angoff method, a more appropriate way of determining the pass mark for the written assessments of the FCP (SA) examination (Part I MCQ test, Part II Objective test and Part II SEQ test)?

In this chapter the performance of the Angoff and Cohen methods of standard setting, for the FCP Part II SEQ tests, have been described and are now compared in a discussion of their respective outcomes.

The discussion is structured in the same order as the results were reported in the preceding text.

7.3.1 Candidates and Cohorts

The criteria to gain entry to sit the Part II exit examination were explained in Chapters 1 and 3. Since these candidates have all i) passed the entry-level Part I MCQ test, and ii) completed almost all their training they are more homogeneous, in terms of examination preparedness, than candidates who write the entry-level Part I MCQ test at the beginning of the 4-year training programme. At the time of conducting this study there was no other information available regarding the profile of candidates, e.g. university at which their primary medical degree was obtained, other postgraduate training completed prior to entering the FCP training programme, university at which specialist training is taking or took place, number of times sitting the examinations, etc. Such additional information may have shed more light on reasons accounting for the differences observed between the entry-level and exit-level examination data. The need for more information when interpreting the FCP examinations data is clear and other authors have reported the use of demographic candidate information to provide a deeper understanding of their assessment data (McManus *et al.* 2014:2-17; McManus *et al.* 2005:2-12; Wakeford *et al.* 2015:2-12).

The number of Part II exit examination candidates has increased over past few years. The examinations investigated between 2001 and 2005 had an average of 24 candidates per FCP (SA) Part II sitting (Burch-papers - *cf.* Chapter 2), while the

average number of candidates sitting the examinations investigated in this study was 62 (median 58). This is most likely due to the single exit examination system introduced by the HPCSA in 2011 (*cf.* Chapter 1). However, because the data collection for the SEQ tests started and ended one sitting earlier than the other two tests investigated in this study, the effect of the single exit examination system on candidate numbers was not as evident as in the SEQ tests.

7.3.2 SEQ test candidate performance data

As explained at the start of this chapter, the way in which the scores were captured in the database used for this study, each test had only ten test items for analysis, while the tests actually contained 20 items. Effectively this reduced the sample size by 50% and may have influenced the calculated reliability of the test data.

The March 2012 SEQ test was the easiest paper with the highest maximum score, 95th percentile, minimum score, mean, Cohen65 pass mark and lowest Cohen65 failure rate. However, its mean was only different to the August 2012 paper and hence it was *not* regarded as an outlier paper compared to the other four SEQ tests. This was in contrast to the February 2013 paper, which was the most difficult paper with the lowest maximum score, 95th percentile, minimum score, mean, Cohen65 pass mark and highest Cohen65 failure rate. This February 2013 paper differed statistically significantly from *all* the other four papers in terms of the performance data and hence it was viewed as an outlier test.

The maximum and minimum scores of the respective SEQ tests were similar, except for the outlier February 2013 paper. In addition, the respective SDs and confidence intervals of the mean test scores were narrow, even for the outlier paper, suggesting that there was little variability between the candidates of the different cohorts. This was probably due to the greater homogeneity of the cohorts and the absence of negative marking.

SEQ test item analysis

The mean ID-value per test cycle, expressed as a percentage, was the same as the mean test score of each cohort (similar methods of calculating the values). Therefore,

the ID-value discussion is similar to that of the mean performance score, which was provided in the preceding text of this section.

The mean DI of the 10 items in each test was *below* 0.20 for *all* the SEQ tests. This suggests that the test items were of a poor quality, i.e. did not clearly separate the good candidates from the poor candidates.

The composite IQI percentage for *all* SEQ test items (n=50) was 14% and across the five cycles the IQI ranged from 0 - 20%. This is yet another finding supporting the suggestion that the SEQ tests were inferior to both the MCQ test items as well as the OT items, in terms of the number of good quality test items contained in the tests. This issue is discussed further in the overall discussion in Chapter 8.

Reliability analysis

The Cronbach's alpha coefficient for the respective SEQ tests ranged from 0.76 – 0.79, with a median of 0.78 and the SEMs of the five test cycles ranged from 3.0 – 3.5, with a median of 3.3. These reliability measures were reasonable and stable across all five cycles, however they need to be higher, given the high-stakes nature of this test (a conjunctive standard apply between the Part II written components). These small SEM values were most likely due to the narrow SDs of the test scores, which in turn were probably caused by the low DI values of the test items. This shows that although a test may be considered reliable, based on Cronbach alpha and SEM values, it does not necessarily make the test results valid (good quality).

7.3.3 SEQ test Angoff data

There were 11-16 subject expert judges involved in each of the five Part II SEQ Angoff standard setting procedures, which is adequate according to the literature previously discussed. On average, the panellists who participated in the Angoff standard setting procedures had 11.7 years of experience as FCP Part II examiners. This means that only experienced clinician judges were involved in the Angoff processes. The content and candidates of the SEQ tests are well known to the judges, since it is a clinically orientated exit-level test, based on Internal Medicine theory and practice, at the level of a specialist physician.

The Angoff pass marks generated by the judges were in an extremely narrow range of 2% (50-52%), which was a third of the range of Cohen65 pass marks (7%). Since model answers were not available at the time when the standard setting procedures took place, it is likely that the panellists tended to provide a score in the middle of the range for each 15-mark item (called central tendency marking). The mean Angoff rating across all 50 items was 51.5% with a SD of 3.1%, which means 95% of the Angoff ratings were between 45.4 to 57.6% (mean ± 1.96xSD). This is illustrated in Figure 7.2, which shows the relationship between the Angoff score, a predicted indicator of test item difficulty and the ID-value percentage, an indicator of test item difficulty based on performance data.

While candidates performed similarly across four of the five test cycles, in terms of means and SDs, the failure rates derived using the Angoff method were widely varying (17 - 71%). Furthermore, the Angoff process failed to identify the February 2013 test as more difficult than the four other tests.

Angoff validity analysis

The relationship between the predicted test item difficulty (Angoff rating %) and actual test item difficulty based on performance data (ID-value %) was evaluated by determining the Pearson correlation coefficient (r) for each individual test as well as for all the test items combined.

For one test cycle (the difficult paper of February 2013), the correlation (r) was the 0.78, which was statistically significant (p =0.008). On deeper analysis of the data, this was one cycle where data on all 20 15-mark items were available and on running a Pearson correlation coefficient (r) across the 20 items in this SEQ paper, the r-value was 0.54, which was statistically significant (p=0.013), and r^2 = 0.29, which was more aligned to some of the OT items' data, but still by far the best for the SEQ papers. The correlations for the four other individual SEQ papers were very weak, ranging from -0.35 (negative) to 0.17 (very weak), with corresponding r^2 values between 0.01 and 0.12 (*cf.* Table 7.4). None of these four correlations were statistically significant (p= 0.33 – 0.84).

The combined Pearson correlation r-value for all the 50 SEQ test items was only 0.10, which was not statistically significant (p=0.483). The corresponding r^2 value of 0.01

meant that only 1% of the relationship could be explained by the data, the remaining 99% was random judgement process error. This data confirm that, in the absence of model answers, the judges were unable to predict the difficulty of the SEQ test items for the borderline exit-level candidate. In these circumstances, the Angoff method of standard setting was clearly inappropriate and invalid.

Angoff reliability analysis

Three measures of reliability were used in this study. The Angoff method had weak reliability over all five cohorts according to the Meskauskas method of evaluating the SD of the Angoff pass marks. The Cohen, Kane and Crooks method showed that the August 2011 judgement was reliable, but the other four cohorts did not meet the standard. The final method, calculating the IRR using the ICC, showed that the ICC point values were acceptable (highlighted green in Table7.4) for three of the five test cycles (August 2011, March 2012, August 2012) (Hallgren 2012:5-6), but the wide CIs made these ICC values less useful, except the August 2011 cohort, which had a CI of only 14%.

Therefore, the reliability of the Angoff method, as used in this study for the high-stakes SEQ test, as judged by the three parameters described, was unacceptable for four of the five test cycles, and only the August 2011 test had an acceptable reliability profile.

7.3.4 SEQ test Cohen data

The 95th percentile range of 11% across five SEQ test cycles, and a resultant Cohen65 pass mark range of 7% can be described as a stable performance of the Cohen method. If the February 2013 test, which was statistically significantly more difficult, is excluded, the Cohen65 pass mark range decreases to only 2% across the other four cycles.

Similar to the OT findings, the validity of the Cohen method did not appear threatened by smaller SEQ test cohorts, as was the case in the MCQ test cohorts. Cohort size and SD had no apparent effect on the 95th percentiles like in the MCQ test, although the SDs were virtually the same across the five SEQ test cohorts. These findings are discussed further in Chapter 8 during the overall discussion. The most likely reason for

this 95th percentile stability is the homogeneity of these exit-level cohorts sitting the SEQ tests, despite the sub-optimal reliability of four out of the five papers.

7.3.5 Comparing the outcomes of the standard setting methods

On reviewing the pass marks and failure rates of the two methods evaluated in this study (Cohen and Angoff methods), as well as the previous fixed 50% pass mark [from Tables 7.3 and Figures 7.3(a-b)]:

Pass marks derived using the Angoff method were consistently higher than those derived using the Cohen method (Cohen65) for all the SEQ test cycles. The Cohen65 pass marks did not overlap with the Angoff pass marks, which were on average 8 percentage points higher. Although they differed from between 6 - 12% on individual cycles, they were also of different 'orders' (similar to the MCQ test and OT), as clearly illustrated by the respective corresponding failure rates which they produced. The most important observation, however, is that the Angoff failure rates ranged from 17 - 71%, as compared to the Cohen65 failure rates ranging from 2 - 9%. If the Angoff method had been used in the CoP it would have led to mass failures in some test cycles.

In the SEQ test papers, the Angoff failure rates narrowly tracked the failure rate of the previous fixed 50% pass mark practice, which was understandable given the close tracking of the Angoff pass marks on the 50% level. It is worth noting that the previously used fixed pass mark of 50% would have yielded a 71% failure rate for the February 2013 test.

7.4 CONCLUSION

This chapter presented the results from the second component of the study, specifically relating to the standard setting of the Part II SEQ test. A focussed discussion was also provided, relating to the findings from this chapter. In addition, this chapter contributes to the overarching discussion presented in the next and final chapter, Chapter 8, where conclusions are drawn and limitations are explained from the study, as well as appropriate recommendations are made.

CHAPTER 8

OVERALL DISCUSSION AND CONCLUSIONS OF STANDARD SETTING FOR SPECIALIST PHYSICIAN EXAMINATIONS IN SOUTH AFRICA

8.1 INTRODUCTION

This study investigated the introduction and implementation of standard setting for specialist physician examinations in South Africa. It was a national study, with data collection conducted over three years (September 2011 – February 2014).

There were three research questions posed in this thesis and each had its related objectives to provide a roadmap towards answering them.

The first research question and objective 1.4.3.1 were addressed by the literature review contained in Chapter 2 and not repeated here. The remaining two research questions, and their related objectives, are restated here to enable ease of reference, before highlighting the major findings of the thesis and providing concluding comments and recommendations of the study.

8.1.1 The first research component – The CoP examiners

Research Question 2: What are the knowledge, attitudes, views and perspectives of CoP examiners regarding standard setting, and do they change with training and exposure to a process of standard setting?

Related objective 1.4.3.2: Determine the knowledge, attitudes, views and perspectives of the CoP examiners about standard setting.

Related objective 1.4.3.3: Design, deliver and evaluate the impact of a seminar dealing with standard setting in the CoP.

The *first research component* addressed Research Question 2 and the two related objectives. This research component explored the process of change (of the 'hearts

and minds') of the CoP examiners regarding the introduction and implementation of standard setting (the Cohen method) for the written components of the FCP (SA) examination. The examiners' self-reported knowledge as well as their attitudes, views and perspectives about standard setting and the Cohen method were investigated using an online questionnaire survey, which was sent out by email to all the CoP examiners at two intervals, February 2013 (T1) and February 2014 (T2), which were respectively 18 months and 30 months after the introduction of standard setting in the CoP. A customised, educational seminar about standard setting in the CoP was also planned, delivered and evaluated between T1 and T2.

Chapter 4 presented the detailed results of the first research component of the study, accompanied by a focussed discussion on the findings.

8.1.2 The second research component – The CoP written assessments

Research question 3: Is using the Cohen method, as compared to Angoff method, a more appropriate way of determining the pass mark for the written assessments of the FCP (SA) examination (Part I MCQ test, Part II Objective test and Part II SEQ test)?

Related objective 1.4.3.4: Determine the performance (pass marks and failure rates) of the Angoff method of standard setting using five cycles of written FCP (SA) examinations data.

Related objective 1.4.3.5: Determine the performance (pass marks and failure rates) of the Cohen method of standard setting using the same five cycles of written FCP (SA) examinations data as in 1.4.3.4.

Related objective 1.4.3.6: Determine the variability of the scores of the top-performing candidates, in order to substantiate the assumption within the Cohen method that the test scores of the top candidates sitting the FCP (SA) written examinations, are stable and, therefore, comparable.

The *second research component* addressed Research Question 3 and the three related objectives. This research component related to the comparison of the performance of the Cohen method versus the Angoff method on the three written components of the

FCP (SA) examination. The methods were directly compared on five test cycles for each of the three written formats included in the study. The analysis of the Part I MCQ test was strengthened by the inclusion, and subsequent analysis, of 20% tracker MCQ test items (n=30), to equate the performance of the heterogeneous Part I cohorts. The tracker items were selected based on their psychometric properties and to be representative of the full MCQ test. The principal assumption of the Cohen method, that top performing candidates are stable from one cycle of a test to the next, was also evaluated.

Chapters 5 - 7 presented the detailed results of the second research component of the study for the Part I MCQ test, the Part II OT and SEQ test respectively, each accompanied by a focussed discussion on the relevant findings in the chapter.

In this final chapter of the thesis an overall discussion is provided, which integrates the relevant findings from the two research components of this study, to address the final, overarching objective (1.4.3.7 – stated below) of Research Question 3, regarding the utility comparison of the two standard setting methods.

Related objective 1.4.3.7: Use the findings of Objectives 1.4.3.2 – 1.4.3.6 to contribute to the evaluation of the utility (as defined in Chapter 2) of the Cohen method, as compared to the Angoff method, for the written FCP (SA) examinations.

The *limitations* of the study are then discussed, followed by the *conclusions* drawn, based on the findings and relevant discussions, used to answer Research Questions 2 and 3. The closing section of this chapter discusses the practical and future research *recommendations* emanating from the study.

The chapter and thesis concludes with a final personal remark from the researcher.

8.2 CHANGE MANAGEMENT IN THE CoP REGARDING STANDARD SETTING

The hearts and minds of CoP examiners regarding standard setting

During the time period from Time 0 (May 2011), when standard setting was introduced and accepted for implementation by the CoP council, up to the first online survey in February 2013 (Time 1), there was considerable 'unfreezing' and change in the CoP with regards to standard setting. Support and adoption for standard setting in general and for the Cohen method was high (mean 71%, *cf.* Table 4.7).

This effectively meant that the diffusion of the standard setting innovation (the Cohen method) reached into the "late majority" phase by Time 1 (T1) according to Rogers's model (2003:281) (*cf.* Figure 8.1). Rogers explains that most of the uncertainty about a new innovation must have been removed for the late majority to accept and adopt it (Rogers 2003:284). Therefore, the levels of certainty, confidence and trust were sufficiently high to enable such a rapid diffusion of standard setting in the CoP.

The most likely explanation for this considerable diffusion was probably the influence and leadership of the CoP change *champions* (Rogers 2003:414), who led and drove the change process from within the organisation. Their important and pivotal role was undeniable, given the radical nature of the changes in a traditional organisation (CMSA) where change is difficult, although in the CoP assessment changes occurred more readily (*cf.* Figure 1.1 in Chapter 1). The champions, with the support of the researcher, seem to have provided the appropriate combination of disconfirmation data leading to survival anxiety, together with enough psychological safety to enable such effective and positive change during this 30 month action research process (Schein 2002:36).

In the period between T1 and T2, uncertainties in the remaining non-adopters were addressed during the seminar on standard setting, in the hope of facilitating additional adoption by the remaining CoP examiners. The Time 2 (T2) survey showed an average reduction of 10% in the number of uncertain non-adopter examiners and a subsequent increase in the average adoption of standard setting from 71% to 83% (*cf.* Table 4.7, 4.8 and Figure 8.1). This indicated that virtually the *entire* late majority had adopted standard setting by T2 (84% is the boundary to the final category – the

"laggards") (Rogers 2003:281). Figure 8.1 illustrates the diffusion of standard setting through the CoP up to February 2014 (T2).

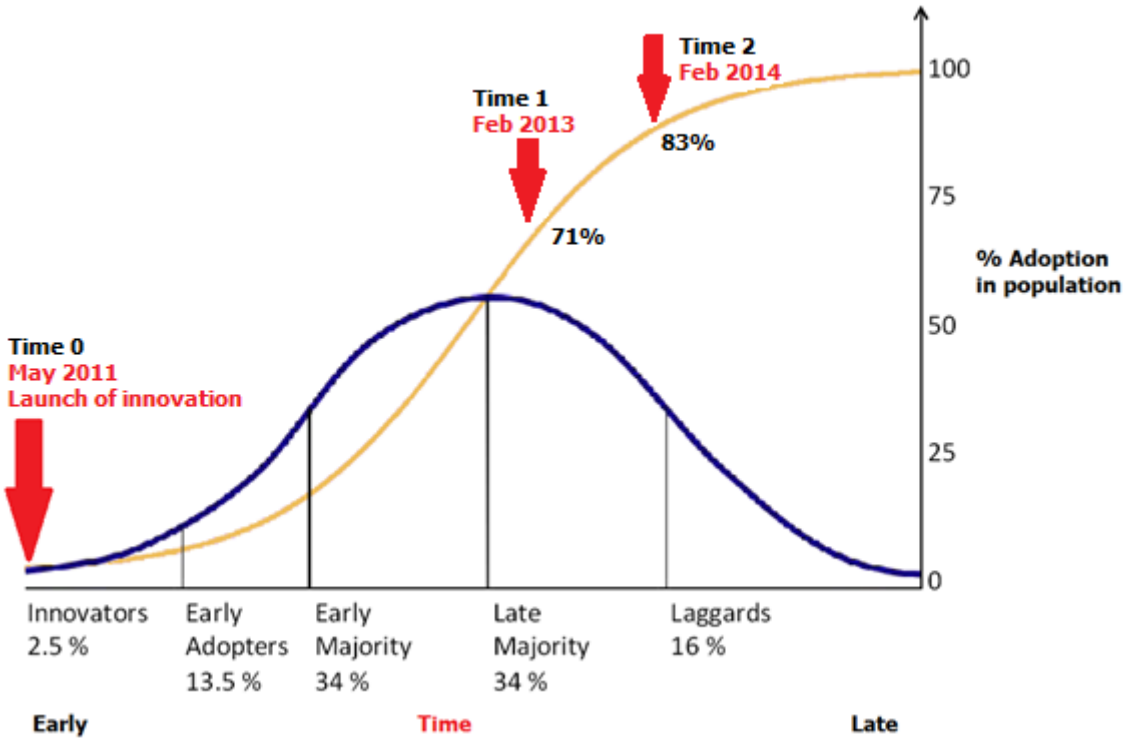


FIGURE 8.1: DIFFUSION OF STANDARD SETTING IN THE CoP BY FEBRUARY 2014
[Adapted from Rogers (2003:281) - Time annotations added]

The T2 survey not only indicated increased mean support, decreased mean uncertainty and similar mean resistance levels to standard setting, but also greater agreement and convergence amongst the CoP examiners. This was evident by the consistent reduction in the standard deviations of the All Item Mean scores in T2 for the "Adopters" and "Non-Adopters" groups (*cf.* Tables 4.7 and 4.8). This finding could be explained by the training (seminar) and additional exposure to standard setting, leading to more certainty and trust in the new system. This was evidence of the development of a new *refreezing* position in the Lewin/Schein model of change (Schein 2002:39), where the change is accepted, widely supported and becomes internalised by the owners of the system, the CoP examiners. The refreezing concept was discussed in Chapter 2, section 2.3.7.3, under the change management section (see section 2.3.7 and Figure 2.10).

The findings of this study suggest that the diffusion of standard setting had not yet penetrated the remaining "laggards" phase. Rogers explains that this final adopter group is categorised as strongly traditionalist, whose "point of reference is the past" (Rogers 2003:284). They need considerably more time to evaluate the new innovation and the relatively short 30 month process was perhaps not yet sufficient for them, or they have concerns and uncertainties that are relevant, but are still unknown and unaddressed by the CoP.

The educational seminar

The impact of the seminar and further exposure of the CoP examiners to standard setting from T1 to T2 was noted in the results of the T2 survey. Statistically significant self-reported increases, from T1 to T2, in the general knowledge about standard setting and the Cohen method as well as the awareness of the implementation of the Cohen method in the OT was most likely the effect of the seminar, since these aspects were directly addressed by the seminar. These results in particular, as well as the results that did not reach statistical significance, but did show a positive improvement in the knowledge, education/training and awareness about standard setting (*cf.* sections 4.2.2 to 4.2.4 in Chapter 4), supported the probable effect and positive impact of the seminar.

However, a clear and definite effect of the seminar was the noted improvement in the attending examiners' (n=24, 44.4%) ability to form an opinion regarding the *utility* aspects of the Cohen and Angoff standard setting methods. Before the seminar, they were essentially unable to rate the methods on the various utility parameters, but all attendees managed to provide ratings afterwards. This indicated an improvement in their knowledge and understanding of the methods and meant that both methods were explicable to an important stakeholder group. The differences in the examiners ratings in the post-seminar evaluation have already been discussed in Chapter 4.

8.3 COMPARING THE COHEN AND ANGOFF METHODS IN THE CoP

As explained in Chapter 2 (sections 2.2.4 – 2.2.5), a critical aspect of quality standard setting is the quality of the assessments used. The accuracy of the pass/fail decisions, derived from the standard setting procedure, is directly influenced by the quality of the assessment data. Therefore, to evaluate the performance of the Cohen and Angoff

methods of standard setting, the quality of the FCP (SA) written tests included in this study had to be determined first.

8.3.1 Quality of the FCP (SA) written tests included in this study

The quality of the written papers included in this study was evaluated by using item difficulty, item discrimination, item quality index and test reliability. All these aspects contribute to the validity of the tests. The blueprints of the tests, which are important sources of validity evidence as well, were not evaluated as part of this study. Therefore, the quality analyses of the papers in this study were evaluated from a purely psychometric perspective. This was a limitation of the study.

Test difficulty

The mean difficulty of the tests included in this study, were fairly similar across the three formats. The median of the five Part I MCQ tests' mean PC-values were 0.55 (range 0.52 – 0.57) and the medians of the Part II OT and SEQ mean ID-values were 0.57 (range 0.54 – 0.62) and 0.56 (range 0.46 – 0.58) respectively. The MCQ tracker mini-tests also had a median of 0.57 (0.56 – 0.59) for the five cycles. Although the medians were similar and of average difficulty and hence acceptable, the range of difficulty varied between the three formats. In the full 150-item MCQ tests, the mean PC-value range was 5% (mini-tests = 2% after rounding), however for the OT it was 8% and for the SEQ it was 12%. The ANOVA analyses confirmed this variability between the three formats, as the MCQ tests had no statistically significant differences between the five mean PC-values, but there were statistically significant differences found between the means of the OT and SEQ tests. This was an important point for the standard setting analysis in this study, because a valid standard setting method must be *sensitive* to changes in test difficulty (Van der Vleuten 2010:175). This point is discussed further later in this section, under the various methods used in this study.

Discrimination ability of the tests

The mean discrimination index (DI) of the test papers were all acceptable, except for the SEQ tests which showed weak DIs of 0.14 – 0.16, with a median of 0.15. The MCQ tests (0.23 – 0.30, median 0.27) and OTs (0.22 – 0.25, median 0.23) were all above the generally accepted 0.20 threshold of acceptable discrimination ability (Downing 2009a:L4747). The low DIs of the SEQ tests were noted as a test quality

concern and the most likely reason for these low DIs were the type of items. They were 30-mark constructed-response items, usually marked without a model answer, resulting in central tendency scores. This point was supported by the narrow standard deviations of the SEQ tests, irrespective of the fluctuations in the difficulty of the papers (*cf.* Table 7.3).

Overall item quality

The overall item quality index (IQI) of the OT (n=150) and MCQ test (n=750) items included in this study, were 63% and 61% respectively. Although this is a relatively new way to express the number of psychometrically sound test items used in an assessment, the IQIs of the five individual OT and MCQ tests were deemed acceptable, since more than half of the items were of good and acceptable quality (IQI>50%). The MCQ tracker items (n=30) had an overall IQI of 90%, which reflected their inherent high psychometric quality for which they were selected. In contrast, the SEQ test items (n=50) had an overall IQI of only 14%, which was concerning. The five individual SEQ tests had IQIs ranging from 0 – 20%, with a median in 20%. This low number of quality items included in each SEQ test meant that the SEQ tests would struggle to discriminate between high and low performing candidates and the difficulty of the tests will potentially fluctuate more. Both these factors raised standard setting concerns for the SEQ test. Criterion-referenced licensing tests should be able to effectively discriminate between competent and incompetent candidates and hence, low quality items in tests would weaken this objective (Downing 2009a:L1651).

Reliability of the tests

The reliability of the tests included in this study was the final psychometric quality parameter used to evaluate the quality of the tests. As discussed in Chapter 2, reliability of test scores is a prerequisite for their validity, but it is not a sufficient source of validity evidence on its own (Axelson & Kreiter 2009:L1003). The individual reliability of the FCP (SA) written components are even more important since conjunctive standards are used, where all tests must be passed independently to pass the overall examination. Compensation between written test formats (in the Part II examination) is not allowed.

The OT and MCQ tests evaluated in this study had excellent reliability, with median Cronbach's alpha coefficients and standard deviation (in brackets) for the five cycles of

0.89 (10.9%) and 0.92 (14.6%) respectively. This was supported by low median SEM values of 3.6% and 4.2% in the OT and MCQ tests respectively. The SEQ tests, by contrast, had lower reliability, with a median Cronbach's alpha coefficient for the five SEQ tests of 0.78 (range 0.76 – 0.79) and median SD of 7.0%, which was below the acceptable 0.80 literature benchmark, as suggested by Hutchinson *et al.* (2002:86) for high-stakes licensing examinations. The median SEM of the SEQ tests were, however, lower than the other two formats at 3.3%, which provides a different picture of the SEQ tests' reliability.

As explained in Chapter 7 (section 7.3.2), one reason for the low alpha coefficients of the SEQ tests was probably due to their low IQI values and resulting narrow standard deviations (SDs), which reduces reliability coefficients. Another possible reason was the small sample size of the SEQ tests, ten items per paper, which will also negatively affect the reliability and its subsequent alpha coefficient calculation.

Therefore, as argued and reported by Tighe *et al.* (2010:1-9), this study also found that the SEM was a better judge of the reliability or reproducibility of the tests, than Cronbach's alpha coefficient. The very high alpha coefficients (>0.90) reported in this study for the Part I MCQ tests were probably an overestimation of their reliability, due to the added CIV and resulting wider SDs introduced by negative marking of the MCQ tests (weaker candidates scoring even lower). Once this effect was corrected, by using the SEM, the reliability impression weakened slightly, when comparing the data with the MRCP (UK) study by Tighe *et al.* (2010:6).

The MRCP (UK) Part I best-of five MCQ papers had 200 items, 50 more than the FCP (SA) Part I MCQ papers, and had a mean alpha coefficient of 0.91, SD of 10.5% and SEM of 3.2% over 16 test cycles reported. In comparison to the present study, the alpha coefficient reliability and SD were lower than for the Part I MCQ tests reported in this study, but so too was the SEM. This highlights the influence of the SD in reliability analysis and the usefulness of the SEM, which can control for fluctuations in the SD. The 50 additional MCQ test items used by the MRCP (UK) should produce higher alpha coefficients than the FCP (SA) MCQ tests (assuming similar IQI of the UK and RSA tests). However, this was not the case, and most likely due to the wider SDs of the FCP (SA) MCQ tests' results, where negative marking was used [not used by the MRCP (UK)]. After correcting for the SDs, by using the SEM, the expected higher reliability

of the MRCP (UK) Part I MCQ tests were demonstrated by their lower mean SEM (3.2%) as compared to the FCP (SA) MCQ tests' (4.2%).

The same effect was noted in the Part II OT and SEQ tests, when comparing it to the Part 2 MRCP (UK) study's data by Tighe *et al.* (2010:6). A direct comparison is not possible in the case of the Part II data, given the differences in the test formats between the two countries (UK= MCQ format, RSA= constructed-response formats). Interestingly, the median SEMs for the five cycles of OT (3.6%) and SEQ tests (3.3%) in the present study did not differ by much from each other and were close to the median SEM of 3.6% of the smallest (in terms of test items) five cycles of MRCP (UK) Part 2 MCQ tests (150 test items). Although this was respectively 5 and 15 times the number of test items than in the OT and SEQ tests, the SEMs were relatively similar.

Unfortunately, neither Tighe *et al.* (2010:1-9), nor a comprehensive literature search produced a single reference for what is deemed an acceptable SEM for a high-stakes licensure assessment. This was surprising given the abundance of references on what is an acceptable reliability coefficient value, which were discussed in Chapter 2 (see section 2.2.4.4). The literature merely argues and explains the merits of the SEM as a more accurate indication of the test's reliability. As Tighe and her colleagues (2010:8) explain: "The most important thing in any high-stakes qualifying examination is the accuracy of the pass mark, which is determined by the SEM (and this, as the simulation has shown, is independent of the reliability and the SD of the candidates)".

Summary of test quality findings

Summarising the findings on the quality of the tests included in this study, it can be deduced that the Part I MCQ tests were of good or acceptable psychometric quality, which would support the measurement objective of this entry-level test. Its measurement objective is to effectively discriminate which candidates, in the heterogeneous cohorts sitting this test, have mastered enough basic medical science knowledge (the reference criterion) to enable them to effectively train to become specialist physicians.

The Part II exit-level written assessments (OT and SEQ test), as explained in Chapter 3 section 3.5.4, are written by homogeneous cohorts of candidates who passed the Part I MCQ test previously, were selected into residency programmes and have completed a

substantial amount of their residency training. The objective of these written assessments is to determine which candidates have reached the desired level of competence (the reference criterion) in terms of specialist-level medical knowledge and application of their knowledge. To support this measurement objective, effective discrimination between the high and low performing candidates is highly desirable. The findings of this study showed that the SEQ test, as compared to the OT, did *not* exhibit any real discrimination ability, although it had a lower median SEM. The OT was reliable *and* could effectively discriminate between candidates. This emphasizes the point made in the literature that although a test might be reliable, it does not guarantee its validity for the desired assessment purpose (Axelson & Kreiter 2009:L1003).

8.3.2 Performance data and the top performing candidates

After evaluating test quality, the next step is to review and compare test performance and subsequently the outcomes of the two standard setting methods. Several authors have argued that tests with similar content, format, difficulty and candidates, leading to similar test performance, should yield similar pass marks, otherwise stakeholders will become sceptical of the pass/fail outcomes (Barman 2008:959; MacCann & Stanley 2010:143). In addition, a point which has been made repeatedly in the literature and this thesis as well, is that where there are indeed significant differences in the difficulty of tests or ability of candidates, the chosen standard setting method must be sensitive to it to remain valid and credible (Bandaranayake 2008:842; Cohen-Schotanus & Van der Vleuten 2010:156; McManus *et al.* 2005:2; Norcini & Shea 1997:48; Van der Vleuten 2010:175).

The performance data of the three written test formats used in the FCP (SA) examination from five consecutive examination cycles, between 2011 – 2014, were included in this study.

The comparability of the entry-level Part I MCQ tests were studied further by using 20% (n=30) common test items (tracker items) in each of the five MCQ tests. The use of equating test items to compare performance of candidates across different sitting of a test is an established strategy in the literature (Bandaranayake 2008:843; McManus *et al.* 2005:2-3; Norcini & Shea 1997:50). The equating strategy used was especially

useful for the Part I MCQ tests, given the lack of available candidate information and the likely heterogeneous nature of the larger, entry-examination cohorts. The findings of this study showed that the five Part I MCQ cohorts' mean performance was the same, once the effect of the CIV introduced to the test scores by the negative marking had been corrected. This was confirmed by the tracker mini-tests (equating process), which were highly representative of the full 150-item MCQ tests and where the mean performances of the five cohorts were also the same, even with the CIV of negative marking. It is therefore reasonable to conclude that the mean difficulty of the MCQ tests and the mean candidate ability for the five cycles were the same.

The large variation in the cohort sizes of the MCQ tests did not play a role in the mean cohort performances. It did seem, however, to play a role in the SD of their respective tests and DIs of the test items. The two smaller cohorts (January 2013 and February 2014) had the lowest SDs, mean DIs and 95th percentile values for the full MCQ test and tracker mini-test. This meant that they also had the lowest Cohen65 pass marks and failure rates of the five MCQ cycles. Therefore, together with conclusion from the preceding paragraph and the fact that the maximum MCQ scores of all the cohorts were higher than any of the 95th percentile values in both the full and tracker tests, it would suggest that the smaller cohorts did not have enough top performing candidates to set Cohen65 pass marks similar to the larger cohorts.

The 95th percentile and resultant Cohen65 pass mark data presented in Chapter 5 were distinctly varied *between* the larger (March 2012, August 2012, June 2013) and the smaller cohort groups, but were remarkably similar *within* the groups. The researcher, after consulting various statisticians, was advised that there is no simple statistical test that could be used to determine if there were statistically significant differences between the Cohen65 pass marks. Therefore, although the data *seem to suggest* a difference between Cohen65 pass marks of the larger and smaller MCQ cohort groups, no deduction could be made about whether the differences noted were in fact statistically significant. It is clear that further research, with more data over a longer time period, is needed to investigate this phenomenon, since it might have been a change occurrence in this study. The importance of conducting further research with more data, over a longer period of time, to clarify an unexpected finding in standard setting research, before jumping to decisive conclusions, was also advocated by a recent UK study (McManus *et al.* 2014:15).

From a fairness perspective, it is helpful to know that the methodology of the Cohen method is such, that the benefit of this validity threat with the Part I MCQ test will always go to the candidates and since this is an entrance examination for specialist physician training, and not an exit, licensing examination, it is probably the fair and appropriate course of action. This threat, although real, is likely to reduce over time given the increasing numbers of candidates sitting the Part I MCQ test, because it is now the only entry examination route.

Interestingly, the possible effect of cohort size on the 95th percentile and Cohen65 pass marks was *not* found in the Part II OT or SEQ test data. Within these exit-level, homogenous cohorts, the SDs, DIs, 95th percentile values and resultant Cohen65 pass marks had no apparent pattern, except that the Cohen65 pass mark followed the rank order of the mean test scores very closely across the five cycles of each test format, which is a hall mark of a credible and valid standard setting method as discussed previously (sensitive to difficulty). The difficulty of the OT and SEQ tests were estimated by using the mean test scores as a marker. The August 2012 OT test was statistically significantly easier than the other OTs and it also had the highest 95th percentile value and resultant Cohen65 pass marks of all the OTs. In the SEQ tests, the February 2013 paper was statistically significantly more difficult than the other SEQ tests and its 95th percentile value and resultant Cohen65 pass mark was distinctly lower than the other SEQ tests.

8.3.3 Acceptable failure rates for the written components of the FCP (SA)

The CoP council recently gave their opinion about what constitutes acceptable failure rates for the Part I and II written components of the FCP (SA) examination. A poll was conducted at the October 2014 CoP council meeting and the outcome was that for the Part I MCQ test, a failure rate of between 20 – 45% was deemed acceptable and for the *combined* written components of the Part II examination, a failure rate of between 20 – 35% of candidates was deemed acceptable (CoP 2014a).

This information was used to judge the acceptability of the failure rates produced by the two standard setting methods evaluated as part of this study.

8.3.4 The performance of the Cohen and Angoff methods

The Cohen method was introduced progressively in the CoP since 2011 for the different written components of the FCP (SA) examination. The model of the Cohen method used in the CoP is the Cohen65 model, which sets the pass mark at 65% of the 95th percentile value of the test scores.

The Angoff method used in this study was the original or Yes/No version and for the reasons explained in this thesis, a purist approach (no reality check, no discussion rounds with re-ratings) was followed. For all the Angoff meetings, the appropriate training and briefings were held with the examiner judging panels and a sufficient number of judges were present at each meeting (ten or more at least), with a mean experience level of 10 years or more involvement in the FCP (SA) examinations.

The outcomes of the previous used fixed 50% pass mark are also discussed in this section, as a reference to the past practice of the CoP. As explained and discussed in Chapter 2 (section 2.2.8.1), the fixed pass mark method is invalid and indefensible from a rationale perspective and the data from this study supports this position.

8.3.4.1 *Validity and reliability of the Cohen method*

The Cohen method is not a predictive method and hence the validity concerns associated with test-centred methods regarding predicting the probability of a borderline candidate's success on a test item is not applicable. The validity of the Cohen method is situated in the rationale of its methodology and its underlying assumptions about the performance data of certain candidates. The most important assumptions of the Cohen method is that the top performing candidates are stable or consistent from cycle to cycle and that they provide the best indication of what was realistically possible to achieve on the given test, since the top performers are also affected by the difficulty of the test. As a result, the methodology of the Cohen method is inherently sensitive to the difficulty of the test, as determined by the top performing candidates.

This sensitivity to test difficulty is a desirable feature of the Cohen method and together with the explicit policy expression of why the Cohen65 model is used in the

CoP (*cf.* Chapter 3, section 3.5.2), it provides evidence of procedural validity (Kane 1994:437) of the method's outcomes. The possible threat to the procedural validity of the Cohen method in the Part I MCQ tests has been discussed already (section 5.3.4). The results from this study did not raise any procedural validity concerns about the use of the Cohen method in the Part II OT or SEQ tests.

The reliability or internal consistency of the Cohen method is perfect, since applying the same model of the method, such as the Cohen65 model, repeatedly on the same performance data will generate identical pass marks. As a result, the internal validity evidence, as described by Kane (1994:445-448), of the Cohen method is strong and robust.

The only remaining evidence needed is external validity evidence (Kane 1994:448-455), which relates to how the outcomes of the Cohen method compare to other methods and if they are viewed as realistic and acceptable to the stakeholders. This validity component is discussed later in this chapter in conjunction with the outcomes of the Angoff method and the acceptable failure rates for the CoP council, which was described previously in this chapter.

8.3.4.2 *Validity and reliability of the Angoff method*

As explained in Chapter 2 (section 2.2.9), the Angoff method is the most used test-centred standard setting method. The basic rationale of its methodology is based on the collective ability of a panel of content experts to judge the difficulty of the individual test items for just-competent or borderline candidates and subsequently predict their probable performance on each test item. Multiple modifications of the Angoff method have been develop over the past 40 years, essentially to improve and strengthen the ability of panels to make these probabilistic predictions more reliably, and importantly, more realistically (valid).

The method commonly described and used in the literature to provide internal validity evidence of the Angoff method is the *validity correlation* between the predicted item difficulty judgements of the panel and the actual empirical item difficulty values (PC-value or ID-value) (Brandon 2004:71; Kane 1994:439; Mee *et al.* 2013:28; Norcini *et al.* 1987:62; Verhoeven *et al.* 1999:836; Verhoeven *et al.* 2002:863).

The higher the Pearson correlation coefficient (r) between the predicted and actual item difficulties, the stronger the internal validity evidence for the Angoff pass mark (Verhoeven *et al.* 2002:863). The literature review by Brandon (2004:71) reported mean correlation coefficients of 0.61 (SD=0.16) for the 29 correlations he reviewed which had a similar purist Angoff methodology to this present study. Taylor (1990:37) defined a high or strong positive correlation coefficient (r) as 0.68 or above. Given the literature's call for high or strong correlations between predicted and actual item difficulties, as evidence of validity for high-stakes assessments (Verhoeven *et al.* 2002:863), the strength of the validity correlations calculated in this study was regarded as critical aspects to evaluate the validity of the Angoff method, as used in this study.

The detailed validity correlations for each test cycle in the three respective test formats are provided in the relevant results chapters (Chapters 5-7). The overall validity correlation data between predicted (Angoff) and actual empirical item difficulties from this study reported statistically significant correlation coefficients of 0.37 (p<0.001) for all the MCQ test items (n=750), 0.37 (p=0.037) for the tracker MCQ items (n=30) and 0.37 (p<0.001) for all the OT test items (n=150). The validity correlation data for all the SEQ test items (n=50) was 0.10 and was not statistically significant (p=0.483). This internal validity evidence from all the tests formats included in this study showed that the Angoff pass marks generated in the CoP were mostly invalid for the high-stakes nature of these tests. This validity deduction was supported by the fact that the Angoff judging panels in this study predicted the mean borderline performance *higher* than the actual mean candidate performance for all five of the MCQ tests, 3/5 OT cycles and 1/5 SEQ tests.

The other source of internal validity evidence determined in this study was the reliability of the Angoff judgements by the different panels. As explained in Chapter 3 section 3.5.6.4, three methods were used to determine the reliability of the Angoff judgements. They were the method described by Cohen et al. (1999:364), the method described by Meskauskas (1986:187-203) and the inter-rater reliability (IRR) measurements. The individual test format reliability outcomes are discussed in the relevant results chapters, but an overall summary is provided here.

The reliability of the Angoff judgements for all the MCQ tests, including the tracker mini-tests returned as unreliable across all three reliability estimates. It was clear that there were no internal consistency within the judging panels for the Part I MCQ tests. The expert clinician judges had a large variance in ratings regarding the difficulty of the items, probably due to the non-clinical nature of the entry-examination and the time since they studied the basic science subjects. In addition, the fact that the answer keys to the items were provided to the judges whilst they made their ratings, may have contributed to the wide ranging views on the difficulty of the MCQ items. As discussed with the literature review in Chapter 2 (section 2.2.9.1), the provision of MCQ answer keys may influence the Angoff ratings of judges, but contrary to the findings by Verheggen et al. (2008:210) who reported an increase in reliability with the provision of answer keys, the findings in this study suggest they had made no impact on the reliability of the MCQ Angoff judgements.

The reliability of the Part II OT and SEQ test Angoff ratings provided somewhat mixed findings from the three methods used to measure Angoff reliability. The method by Meskauskas, which is renowned in the literature for its stringency (Cusimano & Rothman 2003:S90; Yudkowsky & Downing 2008:215), showed similar findings to the literature and the MCQ tests with all of the Angoff ratings for the ten OT and SEQ tests returning as unreliable.

The method by Cohen *et al*. showed that the Angoff ratings for two OTs and one SEQ tests were reliable, the rest were classified as unreliable. Lastly, the IRR, which was determined by the intra-class correlations (ICC), were used to determine the reliability of the Angoff ratings. The ICC values were probably the most robust method of the three since it determines the mean correlations between the different judges on each panel. The median ICC values for the five OT cycles were 0.75, which reflects good reliability (Hallgren 2012:9), but the 95% confidence intervals (CIs) of the individual ICC values were wide (ranging from 0.25 – 0.38) which reduced the confidence in these reliability ratings. A similar finding was made with the ICC values for the SEQ tests' Angoff ratings. Although the first SEQ cycle's Angoff rating had an excellent reliability coefficient of 0.9, with a narrow CI of 0.14, the other four cycles had good ICC values, comparable to the OT cycles, but with even wider CIs. This also reduced the confidence in these SEQ Angoff ICC values.

In summary, it is probably fair to deduce that the clinician judges felt more comfortable to judge the difficulty of the clinically focussed OT (with model answers provided) and SEQ test (no model answers provided) items, which reflected in their Part II Angoff ratings being more reliable than for the Part I MCQ tests. However, the reliability of their Part II Angoff ratings, and subsequent pass marks, must still be classified as *concerning*, given the variable findings from the three methods, the wide CIs of the ICC values and the wide range of ICC values across the five OT and SEQ cycles. As a result, the reliability of the Angoff pass marks, as derived in this study, were unacceptable for the high-stakes written assessments of the FCP (SA), especially since conjunctive standards are used in the CoP, where each written paper must be passed *independently* to progress in the FCP (SA) examination.

The only remaining evidence needed is external validity evidence, which relates to how the outcomes of the Angoff method compare to other methods (the Cohen method in this case) and if they are judged as realistic and acceptable to the stakeholders. This validity component is discussed in conjunction with the outcomes of the Cohen method and the acceptable failure rate guidelines of the CoP council, which was previously described in this chapter.

8.3.4.3 *Performance in the Part I examination (MCQ test)*

A full discussion on the pass marks and resultant failure rates of the Cohen and Angoff methods is provided in Chapter 5. However, a summary of that discussion is provided here, which contributes to the conclusions about the performance of the Cohen and Angoff methods in the CoP.

Figure 5.5(a) shows that although the variability of the Cohen65 and the Angoff pass marks were similar, the Angoff pass marks were consistently higher by 9% or more in each cycle. In contrast, Figure 5.5(b) shows considerably more variation in the resultant Angoff failure rates (65 – 98%, range of 33% and median of 79%) than the Cohen65 failure rates (32 - 51%, range of 19% and median of 43%). The previous indefensible fixed 50% pass mark also returned higher failure rates (49 – 61%, range of 12% and median of 58%) compared to the Cohen method, but they were lower than the Angoff pass marks and showed the least variation of the three sets of failure rates.

The range and median failure rates, derived with the Cohen method, for the five Part I MCQ tests included in this study were slightly lower than the 10-year, UK-graduate data from the comparative UK Internal Medicine written entry-examination, the MRCP (UK) Part 1 MCQ examination, which was recently published (McManus *et al.* 2014:1-19). Its failure rates, using the Angoff/Hofstee hybrid method (2003-2008) and statistical equating method (2008-present), ranged from 35 – 72% (range of 37%), with a mean of 51% (McManus *et al.* 2014:7).

The MCQ tracker mini-test

Given that the tracker mini-test was the same for all cohorts and the subsequent cohorts' mean performance showed no statistically significant difference, the expectation was that the pass marks and failure rates should reflect a similar narrow band. From Figure 5.6(a) it was clear that the pass marks derived by the Cohen method had a fairly stable or narrow band of between 49 – 51%, which included the noted lower pass marks from the two smaller cohorts as explained previously. In contrast, the pass marks generated by the Angoff method showed huge variation from 55 – 81% on the same test. The variation in the tracker mini-test Angoff pass marks were considerably more than for the full MCQ tests which had 120 unique and different items in them. This was further evidence that the Angoff strategy used in this study resulted in invalid pass marks.

Figure 5.6(b) shows the resultant failure rates of the tracker mini-tests from the three methods. This reflected an interesting finding regarding the amount of poorly performing candidates of the different cohorts. The fixed 50% pass mark produced a failure rate range of 12% (43 – 55%), however it was clearly the March 2012 cohort that was responsible for this wide range, due to a larger than expected number of poorly performing students (a longer 'tail'). Excluding the March 2012 cohort reduced the failure rate range to a mere 4% (43 – 47%) across the remaining four cohorts. This finding was also noted with the Cohen method's failure rates across the five cohorts, since the March 2012 and June 2013 cohorts both had Cohen65 pass marks of 55%, but the failure rates differed by 8%, which was similar to the 9% difference of the fixed 50% pass mark method in these two cohorts. The March 2012 cohort seems to have added disproportionately to the failure rate variation of the Cohen method and fixed 50% method.

This was not the case with the Angoff method, where the March 2012 cohort was not the reason for the considerable variation in its reported failure rates. The 81% Angoff pass mark for the January 2013 resulted in a failure rate of 96%, which was a clear outlier. Excluding it reduces the range of Angoff failure rates from 38% to 15%, which is close to the 14% of the Cohen method, if its March 2012 data is also excluded.

Therefore, in the case of the Angoff method, its insensitivity to test difficulty (which is the same in this case) led to an inability to set valid pass marks for the mini-tests, which would have resulted in mass failures due to assessment irregularity on the part of the CoP, whereas with the Cohen method, in contrast, consistent and valid pass marks were set because of its sensitivity to the consistent test difficulty of the mini-tests and the stability of the top performing candidates in the larger cohorts. The concern about the Cohen method's validity for the smaller two cohorts was noted previously, but the benefit of this validity concern will always go to the candidates, which is probably a fair outcome for an entrance examination.

Acceptability of the Part I MCQ failure rates

The only median failure rate that fell within the CoP's acceptable failure rate of 20 – 45% for the Part I MCQ test was the Cohen method (43%). This also compared favourably with the 10-year, UK-graduate failure rate data from the MRCP (UK) Part 1 MCQ examination, which was presented previously in this chapter. The Angoff method yielded considerably higher failure rates which were unacceptable and unrealistic. As explained in Chapter 2, the previous fixed 50% pass mark practice is indefensible from a rationale and methodology perspective and hence not fit for use.

8.3.4.4 *Performance in the written Part II examination (OT and SEQ test)*

A full discussion on the pass marks and resultant failure rates of the Cohen and Angoff methods for the OT and SEQ test formats is provided in Chapter 6 and Chapter 7 respectively. A summary of that discussion, as well as a synthesis and discussion on the effect of using conjunctive standards on the overall failure rates of the written component of the Part II examination is provided. Collectively, these discussions inform the conclusions made about the performance of the Cohen and Angoff methods in the CoP.

Taking into consideration the statistically significantly easier OT (August 2012) and more difficult SEQ test (February 2013), the Cohen method produced stable and valid Cohen65 pass marks in a narrow band of 6% (46 - 52%), that were sensitive to the difficulty of the tests [*cf.* Figures 6.3(a) and 7.3(a)]. The resultant failure rates ranged between 16 – 24% for the OT and 2-9% for the SEQ tests.

The pass marks generated by the Angoff method for the OT and SEQ test cycles reflected no sensitivity to test difficulty, which was not surprising given the evidence presented earlier about the poor validity correlations of the Angoff ratings [*cf.* Figures 6.3(b) and 7.3(b)]. For the OT cycles, the Angoff pass marks ranged from 54 – 66%, which cannot be considered a stable or valid outcome, since the easier OT paper of August 2012 had the second lowest pass mark and its pass mark variation range was double (12%) that of the Cohen method.

In the SEQ tests, where model answers were unavailable and hence the judges only rated the test items by the stated questions, the data suggest that central tendency Angoff ratings occurred for the 30-mark SEQ test items. The February 2013 SEQ test was more difficult than any of the other SEQ tests, and yet it had no clearly discernible effect on the Angoff pass marks of the SEQ tests. This lack of sensitivity to test difficulty of the Angoff method in the SEQ tests, coupled by the central tendency of the Angoff ratings (around 50%), meant that the Angoff pass marks essentially became similar to the previous fixed 50% pass mark and its resultant failure rates [*cf.* Figures 7.3(b)]. The SEQ tests' Angoff failure rates ranged from 18 – 71% (range of 53%) and closely tracked the fixed 50% pass mark's failure rates.

Acceptability of the Part II failure rates

As explained previously in Chapter 3 (section 3.5.5), the CoP uses conjunctive standards and hence a candidate must pass the OT and SEQ test independently to progress to the clinical component of the FCP (SA) Part II examination. In this study, there were only four overlapping cohorts between the OT and SEQ tests (the 2012-2013 cycles), and using the failure rate data from the Cohen and Angoff methods for these four cohorts led to the combined Part II written components failure rate data presented in Table 8.1.

TABLE 8.1: FAILURE RATES FOR THE FCP (SA) PART II WRITTEN COMPONENTS

Cycle (candidates)	Failure rate (%)		
	Cohen method	Angoff method	Fixed 50%
Mar 2012 (57)	21.1	40.4	26.3
Aug 2012 (64)	17.2	46.9	35.9
Feb 2013 (58)	20.7	79.3	70.7
July 2013 (75)	17.3	50.7	28.0

Using the data presented in Table 8.1 in conjunction with the CoP's acceptable failure rates (20 – 35%) for the combined Part II written components, stated earlier in this chapter, it is clear that the Cohen method yielded failure rates that were stable and acceptable to the CoP. The previous fixed 50% pass mark method produced acceptable failure rates for three cohorts. However, similar to the Angoff method, this method was unable to detect the difficult February 2013 paper and would have failed nearly 71% of the cohort, which would have led to large numbers of competent candidates being wrongly excluded (false negative decisions) from the clinical component of the FCP (SA) examination. Therefore, as explained in Chapter 2 (see section 2.2.8.1), the fixed 50% pass mark practice is indefensible from a rationale and methodology perspective. The Angoff method yielded failure rates with a large variability and were unrealistically high, as judged according to the recent CoP's poll and, therefore, unacceptable.

Comparing the current FCP (SA) Part II written data from Table 8.1 with the previous published data from 2001-2005 (*cf.* Chapter 2, Table 2.4), it is interesting to note that the current candidate numbers have already increased by two to three times from the previous data. This trend is likely to continue, as explained before, due to the 2010 HPCSA single exit-examination rule. Interestingly, the failure rates (21.4% on average) from the small 2001-2005 cohorts, using a letter-based grading system (2001-2003) or a fixed 50% pass mark (2004 – 2005), were similar to the current failure rates produced by the Cohen method and not those reported for the 50% pass mark. The reason for this finding is most likely explained by the application of *compensatory* standards in the 2001-2005 data, as compared to the current *conjunctive* approach (where a candidate must pass the written components independently).

In addition, comparing the limited Part II written examinations psychometric data available for the six cohorts between 2001 – 2003, which included all the candidates who wrote the papers (*cf.* Table 2.5), to the data reported in this study, suggests that the performance of the current cohorts on the SEQ test (median score 56.1%) were similar to the older data (mean of 55.4%), but for the OT versus its predecessor, the DI-test, the current candidates perform, on average, 10% better – OT median score was 56.7% and the DI-test mean score was 46.8%. Reasons for this difference are speculative, due to lack of data, but could include differences in test difficulty, due to validity and reliability variations, and/or candidate preparedness and ability.

The failure rate data from the Cohen method also compared favourably to the 9-year, UK-graduate data from the comparative UK Internal Medicine written exit-examination, the MRCP (UK) Part 2 examination, which was recently published (McManus *et al.* 2014:1-19). Its failure rates, using the Angoff/Hofstee hybrid method (2005-2009) and statistical equating method (2010-present), ranged from 14 – 30% (range of 16%), with a mean of 23% (McManus *et al.* 2014:7)

8.3.5 Summary of key findings from the second component of the study

Table 8.2 provides a summary of the most important key findings from the second component of the study, which contributed to the addressing of Research Question 3.

TABLE 8.2: SUMMARY OF KEY FINDINGS FROM THE SECOND COMPONENT OF THE STUDY

Variable	FCP (SA) Part I (entry)*	FCP (SA) Part II (exit)**	
	MCQ test	OT	SEQ test
Cohort nature	Heterogeneous (Unselected)	Homogeneous (Selected, training completed)	
Cohort size range	71 - 151	57 - 79	55 - 75
Overall psychometric quality of test	Good	Good	Poor
Performance data on test cycles	No difference between cycles once CIV of negative marking was removed	One cycle was easier, rest the same	One cycle was harder, rest the same
Cohen method – Cohen65 model			
Pass mark validity	concern raised with smaller cohorts	no concerns, despite small cohorts	
Pass mark reliability	perfect – suitable for high-stakes tests	perfect – suitable for high-stakes tests	
Failure rate acceptability	Realistic and acceptable	Realistic and acceptable	
Angoff method - purist strategy			
Pass mark validity	Major concerns - Invalid	Major concerns - Invalid	
Pass mark reliability	Unreliable – unsuitable for high-stakes tests	Concerns – unsuitable for high-stakes tests	
Failure rate acceptability	Unrealistic and unacceptable	Unrealistic and unacceptable	

* Data from Chapter 5, 8 ** Data from Chapter 6 - 8

8.4 EVALUATION OF THE UTILITY OF THE COHEN AND ANGOFF METHODS

This section synthesizes and integrates the findings and discussion from both components of the study and applies them to the utility parameter framework, Table 2.3, derived during the literature review in Chapter 2, to evaluate the utility of the Cohen method, in comparison to the Angoff method, for the written components of the FCP (SA) examination. This section addresses the final objective (1.4.3.7) to answer Research Question 3.

TABLE 8.3: UTILITY COMPARISON OF THE COHEN AND ANGOFF METHODS USED IN THE CoP

Theme	Category	Parameter	Cohen method	Angoff method
Robust Defensibility	Acceptability	Failure rates	Realistic and acceptable (Table 8.2)	Unrealistic and Unacceptable (Table 8.2)
		Objectivity/ Explicability	Very good (Survey and seminar data)	Acceptable (Survey and seminar data)
	Credibility	Validity	Good, but threat noted in smaller *heterogeneous* cohorts (Table 8.2)	Major concerns – Invalid (Table 8.2)
		Reliability	Perfect (Table 8.2)	Weak – not appropriate for high-stakes testing (Table 8.2)
Responsible Fairness	Defensibility	see above	see above	see above
	Transparency	Accessibility	Good (Survey and seminar data)	Good (Survey and seminar data)
Realistic Practicability	Feasibility & Sustainability	Resources required	Excellent (Survey and seminar data)	Not feasible & sustainable (Survey and seminar data)

Based on the utility comparison presented in Table 8.3, the Cohen method performed considerably better than the Angoff method, in the context of this study. The Angoff method was neither robust (defensible), nor realistic (practicable) and hence would not be suitable for use in high-stakes tests, using the only plausible Angoff strategy for the CoP, as was done in this study, i.e. the purist approach.

8.5 LIMITATIONS OF THE STUDY

The initial May 2011 (Time 0, *cf.* Figure 1.3) level of diffusion of the acceptance of, and support for the use of standard setting through the CoP was not known, since there were no situational analysis data available at Time 0 regarding the CoP examiners' knowledge, attitudes, views and perspectives about standard setting. As a result, the assumption of this study was that there was no diffusion of standard setting at Time 0 in the CoP, which was probably the most likely situation at the time.

The lack of demographic and academic background *candidate information*, as well as the *inability to track the number of attempts* of individual candidates within the various written components of the FCP (SA) examination, due to unique examination numbers being assigned to candidates with each cycle of CMSA examinations, limited the understanding of the composition of the various cohorts, especially for the Part I MCQ test. A better understanding of the effect of cohort size on the Part I MCQ test data in this study was also limited by the small *number of cohort samples* available for analysis at this stage.

The loss of four MCQ tracker items in the Part I MCQ test cycle of January 2013 resulted in a smaller sample size of tracker items, which had a limited negative impact on the reliability coefficient and SEM of the equating mini-test for that cohort.

Psychometric evaluation of the SEQ test data was limited by the lack of detailed performance data for each of the 20 test items contained in each paper. The need to amalgamate the available item data, to produce a uniform dataset for each SEQ test cycle, effectively reduced the sample size of SEQ test items available for analysis by 50% (n=10).

Evaluating the quality of the tests included in this study was limited to a psychometric analysis of the test results. Other validity aspects such as analysing blueprints of the test papers did not form part of this study.

The limitations of cost and time during the biannual CoP council meetings, to conduct the Angoff procedures, necessitated the use of a purist Angoff strategy, which was not ideal, but was all that could realistically be achieved and maintained for the duration of the study. This potentially limited the rigor of the Angoff method used in the CoP as part of this study.

There were no available model answers to provide to the judging panels to use with the Angoff procedure for the SEQ test papers. This was a procedural limitation, given the availability and use of answer keys for the MCQ tests and model answers for the OTs.

8.6 CONCLUSIONS

Given the limitations described in the previous section, a number of conclusions can be drawn from the results and discussion of this study. They address the objectives related to Research Questions 2 and 3, which were stated at the beginning of this chapter.

8.6.1 Conclusions from the first component of the study

Based on the T1 and T2 survey results, as well as the seminar evaluation findings as reported in Chapter 4 and this chapter, the following conclusions can be drawn from the first component of the study.

Conclusion 1

The introduction and implementation of formal standard setting (the Cohen method) for the written components of the FCP (SA) examination, has been widely accepted (adopted) and supported by the CoP and evaluation of the change process showed rapid diffusion of standard setting through the community of CoP examiners. This conclusion addresses objective 1.4.3.2, as stated at the start of this chapter.

Conclusion 2

The knowledge, attitudes, views and perspectives of examiners of the CoP regarding standard setting *did* change with training and exposure to a process of standard setting. This conclusion addresses objective 1.4.3.2, as stated at the start of this chapter.

Conclusion 3

The educational seminar on standard setting, presented between T1 and T2, had a positive impact on the knowledge, attitudes, views and perspectives of examiners of the CoP regarding standard setting. This conclusion addresses objective 1.4.3.3, as stated at the start of this chapter.

Conclusion 4

After attending the seminar, examiners rated the Cohen method higher than the Angoff method in all the utility parameters evaluated. This conclusion addresses objective 1.4.3.7, as stated at the start of this chapter.

8.6.2 Conclusions from the second component of the study

Based on the results presented and discussed in Chapters 5 – 7, as well as the synthesized findings and overall discussion from this chapter, the following conclusions can be drawn for the second component of the study.

Conclusion 5

The pass marks and resultant failure rates yielded by the purist Angoff strategy, as used in this study, were disappointing and unrealistic. This outcome is not too surprising, since Yudkowsky & Downing (2009:L1827) did mention that purist strategies of test-centred methods rarely produce realistic pass marks and failure rates. Although this approach was the only plausible strategy, given the significant resource limitations (human, time, financial) that characterise medical education and training in South Africa, as described by Burch (2007:81-97) and noted in Chapter 2 (section 2.3.3), it was sub-optimal and not fit for purpose. This conclusion addressed objective 1.4.3.4, which is stated at the start of this chapter.

Conclusion 6

The data presented in this study show that the Cohen method was not adversely affected by small cohort sizes in the written tests of the exit-level Part II examination, sat by homogeneous populations of candidates, as previously explained. In contrast, the validity of the Cohen method appears to be threatened by small cohort sizes in heterogeneous populations of examinees, i.e. the candidates sitting the entry-level Part I MCQ test. This conclusion addresses objectives 1.4.3.5 and 1.4.3.6, as stated at the start of this chapter.

Conclusion 7

Given the potential impact of the cohort size on the validity of the Cohen method in heterogeneous examinee cohorts, as described in *Conclusion 6*, it is not possible to comment on the consistency of top performing candidates in the Part I examination

(MCQ test). This conclusion addresses objectives 1.4.3.5 and 1.4.3.6, as stated at the start of this chapter.

Conclusion 8

If results of the statistically significantly easy OT (March 2012) and more difficult SEQ test (February 2013) are removed from the Part II examination data, the remaining data show that the performance of the top candidates, as benchmarked by the 95th percentile, was very consistent. This conclusion addresses objectives 1.4.3.5 and 1.4.3.6, as stated at the start of this chapter.

Conclusion 9

Although noting the possible validity threat described in *Conclusion 6*, the Cohen65 model of the Cohen method yielded valid and reliable pass marks, with acceptable failure rates, for all the written components of the FCP (SA) examination. This conclusion addresses objective 1.4.3.5, which is stated at the start of this chapter.

Conclusion 10

The overall utility of the Cohen method was rated higher than the Angoff method in this study (*cf.* Table 8.3). This conclusion addressed objective 1.4.3.7, which is stated at the start of this chapter.

Conclusions 1 – 3 answered the second research question posed in this thesis. The answer to Research Question 3, based on the data and remaining seven conclusions (*Conclusions 4 – 10)* from this study is: Yes, the Cohen method is a more appropriate way of determining the pass mark for the written assessments of the FCP (SA) examination (Part I MCQ test, Part II Objective test and Part II SEQ test).

8.7 RECOMMENDATIONS FROM THIS STUDY

Based on the outcomes of this study, the following recommendations are made. Firstly the practical recommendations and thereafter the future research recommendations.

8.7.1 Practical recommendations from this study

Recommendation 1

This study supports the introduction and ongoing use of the Cohen65 model of the Cohen method, as a feasible and sustainable method to set the pass marks for the written components of the FCP (SA) examination.

Recommendation 2

The negative marking system used for the Part I MCQ test should be reformed in order to remove the CIV added to the test scores. The system described by Cohen-Schotanus and Van der Vleuten (2010:157) as discussed in this Chapter 2 section 2.2.10 is a plausible alternative.

Recommendation 3

The SEQ test should either be reformed to improve its discrimination ability or replaced, preferably by an assessment instrument that can effectively assess the application of clinical knowledge, such as clinical case-based single best answer MCQs, with three possible options.

Recommendation 4

Given the nature (high-stakes) of the assessments conducted by the CMSA, the Examinations and Credentials Committee (ECC) of the CMSA should consider formulating and publishing an *overarching assessment policy,* including the routine use of standard setting procedures, for use by the member Colleges.

Recommendation 5

The CMSA should consider assigning and using a *single examination number*, unique to each candidate, for the full duration of a candidate's interactions with the CMSA. Effective tracking of candidates will open new research opportunities with regards to assessment, similar to those recently described by McManus and colleagues (McManus *et al.* 2014:12), i.e. evaluating the predictive validity of entry-level examination success on subsequent performance in the exit-level examinations, when changes to assessment practices or standard setting strategies, are implemented. One possibility for the CMSA to consider in this regard is to adopt the readily available and unique number assigned to each doctor upon registration with the HPCSA to practise in South

Africa, the so-called MP-number, as its permanent examination number for each candidate.

Recommendation 6

In addition to using a permanent unique examination number for each candidate, the CMSA should consider gathering appropriate and useful demographic and academic background information on the candidates entering its various Fellowship examinations. This information, tagged with the outcomes of the examinations, could open new research opportunities to the CMSA and provide helpful data and feedback to stakeholders about which factors positively predict for success on CMSA examinations and which factors have no effect on outcomes at all. Naturally, the appropriate procedures surrounding access for researchers to the sensitive and confidential, yet important, data such as assurance of anonymity and other ethical considerations must be in place or need to be developed.

Recommendation 7

Of all the parameters outlined in the standard setting utility framework presented in this thesis, the critical impact of *limited resources* on the implementation and ongoing use of a standard setting method must be recognised and addressed, to ensure that assessment systems are not severely compromised.

Recommendation 8

The use of customised, educational interventions (seminars/workshops) improves knowledge and understanding, reduces uncertainty and thereby, may facilitate the introduction of educational innovations in high-stakes settings.

8.7.2 Future research recommendations from this study

Research recommendation 1

The CoP requires more data, over a prolonged period of time, to sufficiently evaluate the potential impact of cohort size on the consistency of performance data of top performing candidates from entry-level, heterogeneous cohorts. Specifically, to clarify the number of candidates needed to overcome this potential validity threat for the Cohen method.

Research recommendation 2

The findings from this study were limited by a lack of information about the candidates and their number of attempts at the FCP (SA) examinations. It became clear that further research is needed to develop a better understanding about the background of the candidates themselves, the number of attempts at the FCP (SA) examinations and to track individual candidates progress through the FCP (SA) examination process to gather evidence of the predictive validity of the Part I component for the Part II components. This strategy was recently advocated in the literature by McManus and colleagues (2014:14) as an effective way to monitor if changes to standard setting procedures produced results that lead to improved outcomes in subsequent examinations.

Research recommendation 3

Evaluate the usefulness of the standard setting utility parameter framework, as derived in this thesis, in other contexts and when using other assessment instruments, such as performance–based assessment tools, for example an OSCE.

Research recommendation 4

When implementing a major change in medical education practice, such as the introduction of standard setting, it may be helpful to use an action research approach, which enables the stakeholders to monitor and document the outcomes of the process. This may provide useful information for other colleagues, embarking on similar educational change projects.

CONCLUDING PERSONAL REMARK

This research study and thesis reported on the considerable leadership, bravery and progress demonstrated by the CoP (council and examiners) from the inception of this project's journey in May 2011 and which I was privileged to witness and grateful to experience. Their positive engagement with significant assessment changes, relating to how they make pass/fail decisions about the candidates sitting the written components of the FCP (SA) examination, within the context of a conservative testing organisation steeped in 60 years of assessment tradition (CMSA), was truly inspiring and humbling. In addition, the influence of change champions was profound in this project and I experienced first-hand how important they are to enable major change.

I sincerely hope the work presented in this thesis will contribute towards improving the quality assurance and fairness of licensing examinations and certification processes of medical specialists and thereby improve health outcomes in my beloved South Africa. The words of Greg Cizek in 2001 that "it is a dynamic time to be involved in the art and science of standard setting" (Cizek 2001:14), certainly still rings true in my heart today.

Scarpa Schoeman, January 2015

REFERENCES

Agius, S.J., Willis, S.C., McArdle, P.J. & O'Neill, P.A. 2008. Managing change in postgraduate medical education: still unfreezing? *Medical Teacher,* 30(4)**:**e87-e94.

Altman, D.G. 1991. *Practical statistics for medical research (1st edition).* London: Chapman and Hall.

Angoff, W.H. 1971. Scales, norms, and equivalent scores. In Thorndike, R. L. (Ed.) *Educational Measurement* (*2nd edition*). Washington, DC: American Council on Education.

Anonymous author. 1990. Lancet Editorial: Examining the Royal Colleges' examiners. *The Lancet,* 335(8687)**:**443-445.

Axelson, R.D. & Kreiter, C.D. 2009. Chapter 3 - Reliability. In Downing, S. M. & Yudkowsky, R. (Eds) *Assessment in health professions education (ebook)* (*Kindle edition*). New York: Routledge.

Bandaranayake, R.C. 2008. Setting and maintaining standards in multiple choice examinations: AMEE Guide No. 37. *Medical Teacher,* 30(9-10)**:**836-45.

Barman, A. 2008. Standard setting in student assessment: Is a defensible method yet to come? *Annals Academy of Medicine Singapore,* 37(11)**:**957-63.

Ben-David, M.F. 2000. AMEE Guide No. 18: Standard setting in student assessment. *Medical Teacher,* 22(2)**:**120-130.

Berk, R.A. 1986. A consumer's guide to setting performance standards on criterion-referenced tests. *Review of Educational Research,* 56(1)**:**137-172.

Berk, R.A. 1996. Standard setting: The next generation (where few psychometricians have gone before!). *Applied Measurement in Education,* 9(3)**:**215-235.

Beuk, C.H. 1984. A method for reaching a compromise between absolute and relative standards in examinations. *Journal of Educational Measurement,* 21(2)**:**147-152.

Bhandary, S. 2011. Standard Setting in Health Professions Education. *Kathmandu University Medical Journal,* 9(1)**:**3-4.

Boelen, C. 1995. Prospects for change in medical education in the twenty-first century. *Academic Medicine,* 70(7)**:**S21-S28.

Boelen, C., Dharamsi, S. & Gibbs, T. 2012. The social accountability of medical schools and its indicators. *Education for Health,* 25(3)**:**180.

Boelen, C. & Heck, J. (World Health Organization). 1995. *Defining and measuring the social accountability of medical schools.* World Health Organization. Geneva.

Boelen, C. & Woollard, R. 2009. Social accountability and accreditation: a new frontier for educational institutions. *Medical Education,* 43(9)**:**887-894.

Boelen, C. & Woollard, R. 2011. Social accountability: The extra leap to excellence for educational institutions. *Medical Teacher,* 33(8)**:**614-619.

Boulet, J.R., De Champlain, A.F. & McKinley, D.W. 2003. Setting defensible performance standards on OSCEs and standarized patient examinations. *Medical Teacher,* 25(3)**:**245-249.

Boursicot, K.A.M. & Roberts, T.E. 2006. Setting Standards in a Professional Higher Education Course: Defining the Concept of the Minimally Competent Student in Performance ‐ based Assessment at the Level of Graduation from Medical School. *Higher Education Quarterly,* 60(1)**:**74-90.

Boursicot, K.A.M., Roberts, T.E. & Pell, G. 2006. Standard setting for clinical competence at graduation from medical school: a comparison of passing scores across five medical schools. *Advances in Health Sciences Education,* 11(2)**:**173-183.

Bowers, J.J. & Shindoll, R.R. 1989. *A Comparison of the Angoff, Beuk, and Hofstee Methods for setting a passing score*: American College Testing Program.

Brandon, P.R. 2004. Conclusions About Frequently Studied Modified Angoff Standard-Setting Topics. *Applied Measurement in Education,* 17(1)**:**59-88.

Buckendahl, C.W., Smith, R.W., Impara, J.C. & Plake, B.S. 2002. A comparison of Angoff and bookmark standard setting methods. *Educational Measurement,* 39(3)**:**253-263.

Burch, V.C. 2007. *Medical Education in South Africa - assessment practices in a developing country.* PhD thesis. Erasmus University Rotterdam. Published Cape Town, South Africa (ISBN 9780620382236).

Burch, V.C. 2011. Medical education in the 21st century: what would Flexner ask? *Medical Education,* 45(1)**:**22-24.

Burch, V.C. 2014. *Assessment reforms in the CoP from 2003 to 2011*. Personal meeting with Dr FHS Schoeman to explain reforms (April 2014). Cape Town.

Burch, V.C. & Norman, G.R. 2009. Turning words into numbers: establishing an empirical cut score for a letter graded examination. *Medical Teacher,* 31(5)**:**442-446.

Burch, V.C., Norman, G.R., Schmidt, H.G. & Van der Vleuten, C.P.M. 2008. Are specialist certification examinations a reliable measure of physician competence? *Health Science Education,* 13(4)**:**521-533.

Burdick, W.P., Friedman, S.R. & Diserens, D. 2012. Faculty development projects for international health professions educators: Vehicles for institutional change? *Medical Teacher,* 34(1)**:**38-44.

Carson, J.D. 2001. Legal issues in standard setting for licensure and certification. In Cizek, G. J. (Ed.) *Setting performance standards: Concepts, methods, and perspectives.* Mahwah, NJ: Lawrence Erlbaum Associates.

Case, S.M. & Swanson, D.B. 1998. *Constructing written test questions for the basic and clinical sciences (3rd edition).* Philadelphia, PA: National Board of Medical Examiners.

Cetin, S. & Gelbal, S. 2013. A Comparison of Bookmark and Angoff Standard Setting Methods. *Educational Sciences: Theory & Practice,* 13(4).

Chang, L. 1999. Judgmental item analysis of the Nedelsky and Angoff standard-setting methods. *Applied Measurement in Education,* 12(2)**:**151-165.

Chinn, R.N. & Hertz, N.R. 2002. Alternative approaches to standard setting for licensing and certification examinations. *Applied Measurement in Education,* 15(1)**:**1-14.

Christobal, F., Engel, C.E. & Talati, J. 2009. Network: TUFH Position Paper-The Ultimate Challenge? Higher Education for Adapting to Change and Participating in Managing Change. *Education for Health,* 22(3)**:**419.

Cizek, G.J. 2001. *Setting performance standards: Concepts, methods, and perspectives.* Mahwah, NJ: Lawrence Erlbaum Associates.

Cizek, G.J. 2013. *Setting Performance Standards: Theory and Applications.* New York: Routledge.

Cizek, G.J., Bunch, M.B. & Koons, H. 2004. Setting performance standards: contemporary methods. *Educational measurement: issues and practice,* 23(4)**:**31-31.

Clauser, B.E., Harik, P., Margolis, M.J., McManus, I.C., Mollon, J., Chis, L. & Williams, S. 2009a. An empirical examination of the impact of group discussion and examinee performance information on judgments made in the Angoff standard-setting procedure. *Applied Measurement in Education,* 22(1)**:**1-21.

Clauser, B.E., Mee, J., Baldwin, S.G., Margolis, M.J. & Dillon, G.F. 2009b. Judges' Use of Examinee Performance Data in an Angoff Standard‐Setting Exercise for a Medical Licensing Examination: An Experimental Study. *Journal of Educational Measurement,* 46(4)**:**390-407.

Clauser, J.C., Clauser, B.E. & Hambleton, R.K. 2014. Increasing the Validity of Angoff Standards Through Analysis of Judge-Level Internal Consistency. *Applied Measurement in Education,* 27(1)**:**19-30.

CMSA (Colleges of Medicine of South Africa). 2015. *Documents: Examination regulations*. Colleges of Medicine of South Africa. Web Page: http://www.collegemedsa.ac.za/view_document_list.aspx?Keyword=Examination%20Regulations Accessed: 6 January 2015.

Cohen-Schotanus, J. & Van der Vleuten, C.P.M. 2010. A standard setting method with the best performing students as point of reference: Practical and affordable. *Medical Teacher,* 32(2)**:**154-60.

Cohen-Schotanus, J., Van der Vleuten, C.P.M. & Bender, W. 1996. Een betere cesuur bij tentamens [A better standard setting method for written tests]. *Gezond Onderwijs,* 5(0)**:**83-88.

Cohen, A.S., Kane, M.T. & Crooks, T.J. 1999. A generalized examinee-centered method for setting standards on achievement tests. *Applied Measurement in Education,* 12(4)**:**343-366.

Cook, D.A. & Beckman, T.J. 2006. Current concepts in validity and reliability for psychometric instruments: theory and application. *The American journal of medicine,* 119(2)**:**166.e7-e16.

CoP (College of Physicians of South Africa). 2011a. *Minutes of council meeting - May 2011*. Colleges of Medicine of South Africa (CMSA). Johannesburg.

CoP (College of Physicians of South Africa). 2011b. *Regulations for admission to the Fellowship of the College of Physicians of South Africa - FCP (SA)*. Colleges of Medicine of South Africa (CMSA). Johannesburg.

CoP (College of Physicians of South Africa). 2013. *Regulations for admission to the Fellowship of the College of Physicians of South Africa - FCP (SA)*. Colleges of Medicine of South Africa (CMSA). Johannesburg.

CoP (College of Physicians of South Africa). 2014a. *Minutes of council meeting - Oct 2014*. Colleges of Medicine of South Africa (CMSA). Johannesburg.

CoP (College of Physicians of South Africa). 2014b. *Regulations for admission to the Fellowship of the College of Physicians of South Africa - FCP (SA)*. Colleges of Medicine of South Africa (CMSA). Johannesburg.

Crocker, L. & Zieky, M.J. 1995. *Executive Summary on Conference outcomes*. *Joint Conference on Standard Setting for Large-Scale Assessments* Washington, DC.

Cusimano, M.D. 1996. Standard setting in medical education. *Academic Medicine,* 71(Suppl)**:**S112-S120.

Cusimano, M.D. & Rothman, A.I. 2003. The Effect of Incorporating Normative Data into a Criterion-Referenced Standard Setting in Medical Education. *Academic Medicine,* 78(10)**:**S88-S90.

De Gruijter, D.N.M. 1985. Compromise models for establishing examination standards. *Journal of Educational Measurement,* 22(4)**:**263-269.

Dochy, F., Kyndt, E., Baeten, M., Pottier, S. & Veestraeten, M. 2009. The effects of different standard setting methods and the composition of borderline groups: A study within a law curriculum. *Studies in Educational Evaluation,* 35(4)**:**174-182.

Downing, S.M. 2002. Construct‐irrelevant Variance and Flawed Test Questions: Do Multiple‐choice Item‐writing Principles Make Any Difference? *Academic Medicine,* 77(10)**:**S103-S104.

Downing, S.M. 2003a. Guessing on selected‐response examinations. *Medical Education,* 37(8)**:**670-671.

Downing, S.M. 2003b. Validity: on the meaningful interpretation of assessment data. *Medical Education,* 37(9)**:**830-837.

Downing, S.M. 2004. Reliability: on the reproducibility of assessment data. *Medical Education,* 38(9)**:**1006-1012.

Downing, S.M. 2005. The effects of violating standard item writing principles on tests and students: the consequences of using flawed test items on achievement examinations in medical education. *Advances in Health Sciences Education,* 10(2)**:**133-143.

Downing, S.M. 2009a. Chapter 5 - Statistics of Testing. In Downing, S. M. & Yudkowsky, R. (Eds) *Assessment in health professions education (ebook)* (*Kindle edition*). New York: Routledge.

Downing, S.M. 2009b. Chapter 7 - Written Tests. In Downing, S. M. & Yudkowsky, R. (Eds) *Assessment in health professions education (ebook)* (*Kindle edition*). New York: Routledge.

Downing, S.M. & Haladyna, T.M. 2004. Validity threats: overcoming interference with proposed interpretations of assessment data. *Medical Education,* 38(3)**:**327-333.

Downing, S.M. & Haladyna, T.M. 2009. Chapter 2 - Validity and its threats. In Downing, S. M. & Yudkowsky, R. (Eds) *Assessment in health professions education (ebook)* (*Kindle edition*). New York: Routledge.

Downing, S.M., Lieska, N.G. & Raible, M.D. 2003. Establishing Passing Standards for Classroom Achievement Tests in Medical Education: A Comparative Study of Four Methods. *Academic Medicine,* 78(10)**:**S85-S87.

Downing, S.M., Tekian, A. & Yudkowsky, R. 2006. Procedures for establishing defensible absolute passing scores on performance examinations in health professions education. *Teaching and Learning in Medicine,* 18(1)**:**50-57.

Downing, S.M. & Yudkowsky, R. 2009. Chapter 1 - Introduction to Assessment in the Health Professions. In Downing, S. M. & Yudkowsky, R. (Eds) *Assessment in health professions education (ebook)* (*Kindle edition*). New York: Routledge.

Dowton, S.B., Stokes, M., Rawstron, E.J., Pogson, P.R. & Brown, M.A. 2005. Postgraduate medical education: rethinking and integrating a complex landscape. *Medical Journal of Australia,* 182(4)**:**177-80.

Ebel, R.L. 1972. *Essentials of educational measurement (2nd Edition).* Oxford: Prentice-Hall.

Eccles, T. 1994. *Succeeding with Change – Implementation Action Given Strategies.* London: McGraw Hill.

Educational Testing Service 2002. *ETS standards for quality and fairness.* Princeton,NJ.

Epstein, R.M., Cox, M. & Irby, D.M. 2007. Assessment in medical education. *New England Journal of Medicine,* 356(4)**:**387-396.

Ferdous, A.A. & Buckendahl, C.W. 2013. Evaluating Panelists' Standard Setting Perceptions in a Developing Nation. *International Journal of Testing,* 13(1)**:**4-18.

Gale, R. & Grant, J. 1997. AMEE Medical Education Guide No. 10: Managing change in a medical context: guidelines for action. *Medical Teacher,* 19(4)**:**239-249.

GCSA 2010. *Global Consensus for Social Accountability of Medical Schools (GCSA).* East London RSA. Available from www.healthsocialaccountability.org. Accessed on 18 October 2014.

George, S., Haque, M.S. & Oyebode, F. 2006. Standard setting: comparison of two methods. *BMC Medical Education,* 6(1)**:**46.

Gibbs, T. 2011. Sexy words but impotent curricula: Can social accountability be the change agent of the future? *Medical Teacher,* 33(8)**:**605-607.

Gibbs, T. & McLean, M. 2011. Creating equal opportunities: The social accountability of medical education. *Medical Teacher,* 33(8)**:**620-625.

GMC (General Medical Council). 2009. *Tomorrow's Doctors*. General Medical Council. London, UK.

GMC (General Medical Council). 2010. *Standards for curricula and assessment systems*. General Medical Council. London, UK.

Grant, J. 2010. Principles of curriculum design. In Swanwick, T. (Ed.) *Understanding Medical Education: Evidence, Theory, Practice* (*1st edition*). London: Wiley-Blackwell.

Graves, R.R. & Burch, V.C. 2012. SOCKS: launching education innovations on a firm footing. *Medical Education,* 46(11)**:**1122-1123.

Gronlund, N.E. 1998. *Assessment of student achievement (6th edition).* Boston: Allyn & Bacon.

Haladyna, T.M., Downing, S.M. & Rodriguez, M.C. 2002. A review of multiple-choice item-writing guidelines for classroom assessment. *Applied Measurement in Education,* 15(3)**:**309-333.

Hallgren, K.A. 2012. Computing inter-rater reliability for observational data: An overview and tutorial. *Tutorials in quantitative methods for psychology,* 8(1)**:**23.

Hambleton, R.K., Jaeger, R.M., Plake, B.S. & Mills, C. 2000. Setting performance standards on complex educational assessments. *Applied Psychological Measurement,* 24(4)**:**355-366.

Hambleton, R.K. & Pitoniak, M.J. 2006. Setting Performance Standards. In Brennan, R. L. (Ed.) *Educational measurement.* Westport, CT: Praeger Publishers.

Hansen, M.A., Lyon, S.R., Heh, P. & Zigmond, N. 2013. Comparing Panelists' Understanding of Standard Setting Across Multiple Levels of an Alternate Science Assessment. *Applied Measurement in Education,* 26(4)**:**298-318.

Harvill, L.M. 1991. Standard error of measurement. *Educational measurement: issues and practice,* 10(2)**:**181-189.

Hassan, S. 2011. Postgraduate Assessment Rationale of Logical Decisions. *Education in Medicine,* 3(1)**:**e1-e5.

Hays, R.B. 2007. Reforming medical education in the United Kingdom: lessons for Australia and New Zealand. *Medical Journal of Australia,* 187(7)**:**400-403.

Hays, R.B., Gupta, T.S. & Veitch, J. 2008. The practical value of the standard error of measurement in borderline pass/fail decisions. *Medical Education,* 42(8)**:**810-815.

Hays, R.B., Hamlin, G. & Crane, L. 2014. Twelve tips for increasing the defensibility of assessment decisions. *Medical Teacher,* Early Online(0)**:**1-4.

Hess, B., Subhiyah, R.G. & Giordano, C. 2007. Convergence between cluster analysis and the Angoff method for setting minimum passing scores on credentialing examinations. *Evaluation & the health professions,* 30(4)**:**362-375.

Hift, R.J. & Burch, V.C. 2003. Report of symposium on postgraduate assessment within the CMSA. *Transactions: Journal of the Colleges of Medicine of South Africa,* 47(2)**:**75-77.

Hill, D.A., Guinea, A.I. & McCarthy, W.H. 1994. Formative assessment: A student perspective. *Medical Education,* 28(5)**:**394-399.

Hingorjo, M.R. & Jaleel, F. 2012. Analysis of one-best MCQs: the difficulty index, discrimination index and distractor efficiency. *Analysis,* 62(2)**:**142-147.

Hobma, S.O., Ram, P.M., Muijtjens, A.M.M., Grol, R.P.T.M. & Van Der Vleuten, C.P.M. 2004. Setting a standard for performance assessment of doctor– patient communication in general practice. *Medical Education,* 38(12)**:**1244-1252.

Hofstee, W.K.B. 1983. The case for compromise in educational selection and grading. In S.B. Anderson & J.S. Helmick (Ed.) *On educational testing.* San Francisco: Jossey-Bass.

Howell, D.C. 2007. *Statistical Methods for Psychology.* Belmont,CA: Wadsworth.

HPCSA (Health Professions Council of South Africa). 2010. *New requirements for registration of specialists in South Africa*. Health Professions Council of South Africa. Pretoria.

HPCSA (Health Professions Council of South Africa). 2011. *Letter to Medical Deans in RSA: Implementation of the new requirements for the registration of specialists in South Africa*. Health Professions Council of South Africa. Pretoria.

HPCSA & CMSA (Health Professions Council of South Africa & Colleges of Medicine of South Africa). 2014. *Memorandum of understanding between HPCSA and the CMSA.* HPCSA and CMSA. Signed in Pretoria on 17 June 2014.

Hsieh, M. 2013. Comparing Yes/No Angoff and Bookmark Standard Setting Methods in the Context of English Assessment. *Language Assessment Quarterly,* 10(3)**:**331-350.

Hudson (Jnr.), J.P. & Campion, J.E. 1994. Hindsight Bias in an Application of the Angoff Method for Setting Cutoff Scores. *Journal of Applied Pshychology,* 79(6)**:**860-865.

Hurtz, G.M. & Auerbach, M.A. 2003. A meta-analysis of the effects of modifications to the Angoff method on cutoff scores and judgment consensus. *Educational and Psychological Measurement,* 63(4)**:**584-601.

Hutchinson, L., Aitken, P. & Hayes, T. 2002. Are medical postgraduate certification processes valid? A systematic review of the published evidence. *Medical Education,* 36(1)**:**73-91.

IBM Corp. 2013. *IBM SPSS Statistics for Windows.* Version 22.0. Armonk, NY: IBM corporation.

Impara, J.C. & Plake, B.S. 1997. Standard setting: An alternative approach. *Journal of Educational Measurement,* 34(4)**:**353-366.

Impara, J.C. & Plake, B.S. 1998. Teachers' ability to estimate item difficulty: A test of the assumptions in the Angoff standard setting method. *Journal of Educational Measurement,* 35(1)**:**69-81.

Iobst, W.F., Sherbino, J., Ten Cate, O., Richardson, D.L., Dath, D., Swing, S.R., Harris, P., Mungroo, R., Holmboe, E.S. & Frank, J.R. 2010. Competency-based medical education in postgraduate medical education. *Medical Teacher,* 32(8)**:**651-656.

Jaeger, R.M. 1989. Certification of student competence. In Linn, R. L. (Ed.) *Educational Measurement* (*3rd edition*). New York, NY: MacMillian.

Jolly, B. 2010. Written examinations. In Swanwick, T. (Ed.) *Understanding Medical Education: Evidence, Theory, Practice* (*1st edition*). London: Wiley-Blackwell.

Joubert, G. & Katzenellenbogen, J.M. 2007. Population and sampling. In Joubert, G., Ehrlich, R., Katzenellenbogen, J. M. & Karim, S. A. (Eds) *Epidemiology: A Research Manual for South Africa* (*2nd edition*). Cape Town: Oxford University Press Southern Africa.

Kane, M. 1994. Validating the performance standards associated with passing scores. *Review of Educational Research,* 64(3)**:**425-461.

Kane, M. 1998. Choosing Between Examinee-Centered and Test-Centered Standard-Setting Methods. *Educational Assessment,* 5(3)**:**129-145.

Karantonis, A. & Sireci, S.G. 2006. The bookmark standard - setting method: A literature review. *Educational measurement: issues and practice,* 25(1)**:**4-12.

Katz, I.R. & Tannenbaum, R.J. 2014. Comparison of Web-based and Face-to-face Standard Setting using the Angoff Method. *Journal of Applied Testing Technology,* 15(1)**:**1-17.

Katzenellenbogen, J.M. & Joubert, G. 2007. Data collection and measurement. In Joubert, G., Ehrlich, R., Katzenellenbogen, J. M. & Karim, S. A. (Eds) *Epidemiology: A Research Manual for South Africa* (*2nd edition*). Cape Town: Oxford University Press Southern Africa.

Kaufman, D.M., Mann, K.V., Muijtjens, M.M. & Van der Vleuten, C.P.M. 2000. A comparison of standard-setting procudures for an OSCE in Undergraduate Medical Education. *Academic Medicine,* 75(3)**:**267-271.

Kellow, J.T. & Willson, V.L. 2008. Setting standards and establishing cut scores on criterionreferenced assessments. In Osborne, J. W. (Ed.) *Best practices in quantitative methods.* Thousand Oaks, CA: Sage Publications.

Kent, A. 2003. General Comment on the CMSA Educational Symposium - May 2003. *Transactions: Journal of the Colleges of Medicine of South Africa,* 47(2)**:**78-79.

Kwizera, E.N. & Iputo, J.E. 2011. Addressing social responsibility in medical education: The African way. *Medical Teacher,* 33(8)**:**649-653.

Langley, G.J., Moen, R., Nolan, K.M., Nolan, T.W., Norman, C.L. & Provost, L.P. 2009. *The improvement guide: a practical approach to enhancing organizational performance (2nd edition).* Hoboken, NJ: John Wiley & Sons.

Lewin, K. 1946. Action research and minority problems. *Journal of social issues,* 2(4)**:**34-46.

Lewin, K. 1952. Group Desicion and Social Change. In Swanson, G. E., Newcomb, T. M. & Hartley, E. L. (Eds) *Readings in social psychology* (*Revised edition*). New York: Holt.

Lewis, D.M., Mitzel, H.C. & Green, D.R. 1996. *Standard setting: A bookmark approach*. *DR Green (Chair), IRT-based standard setting procedures utilizing behavioral anchoring. Symposium conducted at the Council of Chief State School Officers National Conference on Large-Scale Assessment. Phoenix, AZ.*

Lieff, S. & Albert, M. 2012. What do we do? Practices and learning strategies of medical education leaders. *Medical Teacher,* 34(4)**:**312-319.

Lindgren, S., Ahn, D., Alwan, I.A., Cassimatis, E.G., Jacobs, M.E., Karle, H., Kloiber, O., Van Lerberghe, W., Patricio, M.F., Pulido, P., Sood, R. & Weggemans, M. (World Federation for Medical Education). 2012. *WFME Global Standards for Quality improvement in basic medical education.* Copenhagen.

Lindgren, S. & Karle, H. 2011. Social accountability of medical education: aspects on global accreditation. *Medical Teacher,* 33(8)**:**667-672.

Livingston, S.A. & Zieky, M.J. 1982. *Passing Scores: A Manual for Setting Standards of Performance on Educational and Occupational Tests.* Princeton, NJ: Educational Testing Service.

Livingston, S.A. & Zieky, M.J. 1989. A comparative study of standard-setting methods. *Applied Measurement in Education,* 2(2)**:**121-141.

Lynch, D.C., Surdyk, P.M. & Eiser, A.R. 2004. Assessing professionalism: a review of the literature. *Medical Teacher,* 26(4)**:**366-373.

Lypson, M.L., Downing, S.M., Gruppen, L.D. & Yudkowsky, R. 2013. Applying the Bookmark method to medical education: Standard setting for an aseptic technique station. *Medical Teacher,* 35(7)**:**581-585.

MacCann, R.G. & Stanley, G. 2010. Extending participation in standard setting: an online judging proposal. *Educational Assessment, Evaluation and Accountability,* 22(2)**:**139-157.

MacFarlane, F., Gantley, M. & Murray, E. 2002. The CeMENT project: a case study in change management. *Medical Teacher,* 24(3)**:**320-326.

Maykut, P. & Morehouse, R. 1994. *Beginning Qualitative Research: A Philosophic and Practical Guide.* London: The Falmer Press.

McGaghie, W.C., Butter, J. & Kaye, M. 2009. Chapter 8 - Observational Assessment. In Downing, S. M. & Yudkowsky, R. (Eds) *Assessment in health professions education (ebook)* (*Kindle edition*). New York: Routledge.

McKinley, D.W. & Norcini, J.J. 2014. How to set standards on performance-based examinations: AMEE Guide No. 85. *Medical Teacher,* 36(2)**:**97-110.

McManus, I.C., Chis, L., Fox, R., Waller, D. & Tang, P. 2014. Implementing statistical equating for MRCP (UK) parts 1 and 2. *BMC Medical Education,* 14(1)**:**1-19.

McManus, I.C., Mollon, J., Duke, O.L. & Vale, J.A. 2005. Changes in standard of candidates taking the MRCP (UK) Part 1 examination, 1985 to 2002: analysis of marker questions. *BMC medicine,* 3(1)**:**13.

McManus, I.C., Mooney‐Somers, J., Dacre, J.E. & Vale, J.A. 2003. Reliability of the MRCP (UK) Part I examination, 1984‐2001. *Medical Education,* 37(7)**:**609-611.

McNiff, J. 2013. *Action research: Principles and practice (3rd Edition).* New York: Routledge.

McNiff, J. & Whitehead, J. 2010. *You and your action research project (3rd Edition).* New York: Routledge.

Mee, J., Clauser, B.E. & Margolis, M.J. 2013. The Impact of Process Instructions on Judges' Use of Examinee Performance Data in Angoff Standard Setting Exercises. *Educational measurement: issues and practice,* 32(3)**:**27-35.

Mehta, C.R. & Patel, N.R. 1989. *IBM SPSS exact tests*. SPSS Inc., an IBM company. Camebridge, Massachusetts.

Mennin, S.P. & Kaufman, A. 1989. The change process and medical education. *Medical Teacher,* 11(1)**:**9-16.

Meskauskas, J.A. 1986. Setting Standards for Credentialing Examinations An Update. *Evaluation & the health professions,* 9(2)**:**187-203.

Meskauskas, J.A. & Webster, G.D. 1975. The American Board of Internal Medicine Recertification Examination: Process and Results. *Annals of Internal Medicine,* 82(4)**:**577-581.

Messick, S. 1990. *Validity of test interpretation and use.* Educational Testing Service. Princeton, N.J.

Microsoft. 2007. *Microsoft Excel.* Redmond, Washington: Microsoft.

Miller, G.E. 1990. The assessment of clinical skills/competence/performance. *Academic Medicine,* 65(9)**:**S63-S67.

Morroni, C. & Myer, L. 2007. Study design. In Joubert, G., Ehrlich, R., Katzenellenbogen, J. M. & Karim, S. A. (Eds) *Epidemiology: A Research Manual for South Africa* (*2nd edition*). Cape Town: Oxford University Press Southern Africa.

Myer, L. & Karim, S.A. 2007. Precision and validity in epidemiological studies: Error, bias and counfounding. In Joubert, G., Ehrlich, R., Katzenellenbogen, J. M. & Karim, S. A. (Eds) *Epidemiology: A Research Manual for South Africa* (*2nd edition*). Cape Town: Oxford University Press Southern Africa.

Nedelsky, L. 1954. Absolute grading standards for objective tests. *Educational and Psychological Measurement,* 14(1)**:**3-19.

Norcini, J.J. 2003. Setting standards for educational tests. *Medical Education,* 37(5)**:**464-469.

Norcini, J.J., Anderson, B., Bollela, V., Burch, V.C., Costa, M.J., Duvivier, R., Galbraith, R., Hays, R.B., Kent, A., Perrott, V. & Roberts, T.E. 2011. Criteria for good assessment: Consensus statement and recommendations from the Ottawa 2010 Conference. *Medical Teacher,* 33(3)**:**206-14.

Norcini, J.J. & Banda, S.S. 2011. Increasing the quality and capacity of education: the challenge for the 21st century. *Medical Education,* 45(1)**:**81-86.

Norcini, J.J. & Guille, R. 2002. Combining Tests and Setting Standards. In Norman, G. R., Vleuten, C. P. M., Newble, D. I., Dolmans, D. H. J. M., Mann, K. V., Rothman, A. I. & Curry, L. (Eds) *International Handbook of Research in Medical Education.* Dordrecht: Kluwer Academic Publishers.

Norcini, J.J., Lipner, R.S., Langdon, L.O. & Strecker, C.A. 1987. A Comparison of Three Variations on a Standard - Setting Method. *Journal of Educational Measurement,* 24(1)**:**56-64.

Norcini, J.J. & McKinley, D.W. 2007. Assessment methods in medical education. *Teaching and teacher education,* 23(3)**:**239-250.

Norcini, J.J. & Shea, J.A. 1997. The credibility and comparability of standards. *Applied Measurement in Education,* 10(1)**:**39-59.

Norcini, J.J., Shea, J.A. & Kanya, D.T. 1988. The effect of various factors on standard setting. *Journal of Educational Measurement,* 25(1)**:**57-65.

Norman, G. & Eva, K.W. 2010. Quantitative research methods in medical education. In Swanwick, T. (Ed.) *Understanding Medical Education: Evidence, Theory, Practice* (*1st edition*). London: Wiley-Blackwell.

Parmar, D., Shah, C. & Parmar, R. 2014. Study of standard setting in constructed response type written examination. *Int J Med Sci Public Health,* 3(9)**:**1-5.

Pell, G., Fuller, R., Homer, M. & Roberts, T.E. 2010. How to measure the quality of the OSCE: A review of metrics-AMEE guide no. 49. *Medical Teacher,* 32(10)**:**802-811.

Petersen, S.A. 2013. *Standard setting for Assessments*. *Workshop: Assessment in Medical Edcuation.* Bloemforntein, RSA.

Peterson, C.H., Schulz, E.M. & Engelhard (Jnr.), G. 2011. Reliability and Validity of Bookmark ‐ Based Methods for Standard Setting: Comparisons to Angoff ‐ Based Methods in the National Assessment of Educational Progress. *Educational measurement: issues and practice,* 30(2)**:**3-14.

Pitoniak, M.J. & Yeld, N. 2013. Standard Setting Lessons Learned in the South African Context: Implications for International Implementation. *International Journal of Testing,* 13(1)**:**19-31.

Plake, B.S. 1998. Setting Performance Standards for Professional Licensure and Certification, Applied Measurement in Education. *Applied Measurement in Education,* 11(1)**:**65-80.

Prideaux, D. 2004. Managing change in medical education. *Triannual newsletter produced by the centre for development of teaching and learning,* 8(3)**:**1-16.

Reznick, R.K., Blackmore, D., Dauphinee, W.D., Rothman, A.I. & Smee, S. 1996. Large-scale high-stakes testing with an OSCE: report from the Medical Council of Canada. *Academic Medicine,* 71(Suppl.)**:**S19-21.

Ricker, K.L. 2006. Setting cut-scores: A critical review of the Angoff and modified Angoff methods. *Alberta journal of educational research,* 52(1)**:**53-64.

Riel, M. 2010. *Understanding action research.* Center For Collaborative Action Research, Pepperdine University (Last revision Sep 2013) Accessed on 28 Aug 2014 from http://cadres.pepperdine.edu/ccar/define.html.

Ritter, P., Lorig, K., Laurent, D. & Matthews, K. 2004. Internet versus mailed questionnaires: A randomized comparison. *Journal of Medical Internet Research,* 6(3)**:**e29-e37.

Roberts, C., Newble, D., Jolly, B., Reed, M. & Hampton, K. 2006. Assuring the quality of high-stakes undergraduate assessments of clinical competence. *Medical Teacher,* 28(6)**:**535-543.

Rodriguez, M.C. 2005. Three options are optimal for multiple - choice items: A meta - analysis of 80 years of research. *Educational measurement: issues and practice,* 24(2)**:**3-13.

Rogers, E.M. 2003. *Diffusion of innovations (5th edition).* New York: Simon and Schuster.

Rolfe, G. 2006. Validity, trustworthiness and rigour: quality and the idea of qualitative research. *Journal of Advanced Nursing,* 53(3)**:**304-310.

RSA (Republic of South Africa). 2008. *No. 67 of 2008: National Qualifications Framework Act, 2008*. Cape Town: Government Printer.

RSA (Republic of South Africa). 2009. *Health Professions Act 56 of 1974*. Pretoria: Government Printer.

RSA DBE (Republic of South Africa. Department of Basic Education). 2014. *Origin of the DBE and DHET*. Department of Basic Education. Web Page: http://www.education.gov.za/TheDBE/AboutDBE/tabid/435/Default.aspx Accessed: 4 March 2014.

RSA DoHET (Republic of South Africa. Department of Higher Education and Training). 2013. *NQF Act (67/2008): Amendment to the Determination of the Sub-frameworks that comprise the National Qualifications Framework*. Pretoria: Government Printer.

Sadler, D.R. 2005. Interpretations of criteria - based assessment and grading in higher education. *Assessment & Evaluation in Higher Education,* 30(2)**:**175-194.

SAQA (South African Qualifications Authority). 2000a. *The National Qualifications framework and Quality Assurance*. South African Qualifications Authority. Pretoria, South Africa.

SAQA (South African Qualifications Authority). 2000b. *The National Qualifications framework and the Standards Setting*. South African Qualifications Authority. Pretoria, South Africa.

SAQA (South African Qualifications Authority). 2001. *Criteria and Guidelines for Assessment of NQF Registered Unit Standards and Qualifications*. South African Qualifications Authority. Pretoria, South Africa.

SAQA (South African Qualifications Authority). 2012. *South African Qualifications Authority Professional Body Registration: HPCSA - Health Professions Council of South Africa* SAQA. Web Page: http://pbdesig.saqa.org.za/viewProfessionalBody.php?id=692 Accessed: 31 October 2014.

Schein, E.H. 1996. Kurt Lewin's change theory in the field and in the classroom: Notes toward a model of managed learning. *Systems practice,* 9(1)**:**27-47.

Schein, E.H. 2002. Models and tools for stability and change in human systems. *Reflections,* 4(2)**:**34-46.

Schoeman, S. 2011. Setting standards in health sciences education-a wake-up call. *African Journal of Health Professions Education,* 3(1)**:**2.

Schuwirth, L.W.T. & Van der Vleuten, C.P.M. 2010. How to design a useful test: Principles of assessment. In Swanwick, T. (Ed.) *Understanding Medical Education: Evidence, Theory, Practice* (*1st edition*). London: Wiley-Blackwell.

Schuwirth, L.W.T. & van der Vleuten, C.P.M. 2011. General overview of the theories used in assessment: AMEE Guide No. 57. *Med Teach,* 33(10)**:**783-97.

Schön, D.A. 1963. Champions for radical new inventions. *Harvard business review,* 41(2)**:**77-86.

Searle, J. 2000. Defining competency–the role of standard setting. *Medical Education,* 34(5)**:**363-366.

Shulruf, B., Turner, R., Poole, P. & Wilkinson, T. 2013. The Objective Borderline method (OBM): a probability-based model for setting up an objective pass/fail cut-off score in medical programme assessments. *Advances in Health Sciences Education,* 18(2)**:**231-244.

Shumway, J.M. & Harden, R.M. 2003. AMEE Guide No. 25: The assessment of learning outcomes for the competent and reflective physician. *Medical Teacher,* 25(6)**:**569-84.

Sim, S. & Rasiah, R.I. 2006. Relationship between item difficulty and discrimination indices in true/false-type multiple choice questions of a para-clinical multidisciplinary paper. *Annals-Academy of Medicine Singapore,* 35(2)**:**67.

Smith, R.W., Davis-Becker, S.L. & O'Leary, L.S. 2014. Combining the best of Two Standard Setting Methods: the Ordered Item Booklet Angoff. *Journal of Applied Testing Technology,* 15(1)**:**18-26.

Southgate, L., Hays, R.B., Norcini, J.J., Mulholland, H., Ayers, B., Woolliscroft, J., Cusimano, M.D., McAvoy, P., Ainsworth, M., Haist, S. & Campbell, M. 2001. Setting performance standards for medical practice: a theoretical framework. *Medical Education,* 35(5)**:**474-481.

Spears, K. & Wilson, M. 2010. "I Don't Know" and Multiple Choice Analysis of Pre-and Post-Tests. *Journal of Extension,* 48(6)**:**1-8.

Statistics Solutions 2014. *McNemar, Marginal Homogeneity, Sign, Wilcoxon Tests*. From: Statistics Solutions. Web Page: http://www.statisticssolutions.com/mcnemar-marginal-homogeneity-sign-wilcoxon-tests Accessed: 9 September 2014.

Stern, D.T., Friedman Ben-David, M., De Champlain, A.D., Hodges, B., Wojtczak, A. & Schwarz, M.R. 2005. Ensuring global standards for medical graduates: a pilot study of international standard-setting. *Medical Teacher,* 27(3)**:**207-213.

SurveyMonkey Inc. 2013. *SurveyMoney.* Palo Alto, California, USA: SurveyMonkey.

SurveyMonkey Inc. 2014. *How many respondents do I need?* SurveyMonkey. Web Page: http://help.surveymonkey.com/articles/en_US/kb/How-many-respondents-do-I-need Accessed: 12 June 2014.

Swanwick, T. & Buckley, G. 2010. Introduction: Understanding Medical Education. In Swanwick, T. (Ed.) *Understanding Medical Education: Evidence, Theory, Practice* (*1st edition*). London: Wiley-Blackwell.

Tavakol, M. & Dennick, R. 2011a. Making sense of Cronbach's alpha. *International Journal of Medical Education,* 2(0)**:**53-55.

Tavakol, M. & Dennick, R. 2011b. Post-examination analysis of objective tests. *Medical Teacher,* 33(6)**:**447-458.

Tavakol, M. & Dennick, R. 2013. Modelling the Hofstee method reveals problems. *Medical Teacher,* Early Online(0)**:**1.

Tavakol, M. & Sandars, J. 2014a. Quantitative and qualitative methods in medical education research: AMEE Guide No 90: Part I. *Medical Teacher,* 36(9)**:**746-756.

Tavakol, M. & Sandars, J. 2014b. Quantitative and qualitative methods in medical education research: AMEE Guide No 90: Part II. *Medical Teacher,* 36(10)**:**838-848.

Taylor, C.A. 2011. Development of a modified Cohen method of standard setting. *Medical Teacher,* 33(12)**:**e678-e682.

Taylor, R. 1990. Interpretation of the correlation coefficient: a basic review. *Journal of diagnostic medical sonography,* 6(1)**:**35-39.

Tighe, J., McManus, I.C., Dewhurst, N.G., Chis, L. & Mucklow, J. 2010. The standard error of measurement is a more appropriate measure of quality for postgraduate medical assessments than is reliability: an analysis of MRCP (UK) examinations. *BMC Medical Education,* 10(1)**:**40.

Todd, I., Burr, S., Whittle, J. & Fairclough, L. 2014. Modifying the Hofstee method may overcome problems. *Medical Teacher,* Early Online(0)**:**1.

Tormey, W. 2014. Education, learning and assessment: current trends and best practice for medical educators. *Irish journal of medical science,* Published online - Feb 2014(0)**:**1-12.

Trafford, V.N. & Leshem, S. 2008. *Stepping stones to achieving your doctorate: by focussing on your viva from the start.* Maidenhead,UK: Open University Press.

Twycross, A. 2005. Validity and reliability - What's it all about? *Paediatric nursing,* 17(1)**:**36.

Van der Vleuten, C.P.M. 1996. The assessment of professional competence: developments, research and practical implications. *Advances in Health Sciences Education. Theory and Practice.,* 1(1)**:**41-67.

Van der Vleuten, C.P.M. 2000. Validity of final examinations in undergraduate medical training. *British Medical Journal,* 321(7270)**:**1217-1219.

Van der Vleuten, C.P.M. 2010. Setting and maintaining standards in multiple choice examinations: Guide supplement 37.1-Viewpoint. *Medical Teacher,* 32(2)**:**174-176.

Van der Vleuten, C.P.M. & Schuwirth, L.W.T. 2005. Assessing professional competence: from methods to programmes. *Medical Education,* 39(3)**:**309-317.

Verheggen, M.M., Muijtjens, A.M.M., Van Os, J. & Schuwirth, L.W.T. 2008. Is an Angoff standard an indication of minimal competence of examinees or of judges? *Advances in Health Sciences Education,* 13(2)**:**203-211.

Verhoeven, B.H., Van der Steeg, A.F.W., Scherpbier, A.J.J.A., Muijtjens, A.M.M., Verwijnen, G.M. & Van der Vleuten, C.P.M. 1999. Reliability and credibility of an Angoff standard setting procedure in progress testing using recent graduates as judges. *Medical Education,* 33(11)**:**832-837.

Verhoeven, B.H., Verwijnen, G.M., Muijtjens, A.M.M., Scherpbier, A.J.J.A. & Van der Vleuten, C.P.M. 2002. Panel expertise for an Angoff standard setting procedure in progress testing: item writers compared to recently graduated students. *Medical Education,* 36(9)**:**860-867.

Wakeford, R., Denney, M., Ludka-Stempien, K., Dacre, J.E. & McManus, I.C. 2015. Cross-comparison of MRCGP & MRCP (UK) in a database linkage study of 2,284 candidates taking both examinations: assessment of validity and differential performance by ethnicity. *BMC Medical Education,* 15(1)**:**1-12.

Ware, J. & Vik, T. 2009. Quality assurance of item writing: during the introduction of multiple choice questions in medicine for high stakes examinations. *Medical Teacher,* 31(3)**:**238-243.

Wass, V., Van der Vleuten, C.P.M., Shatzer, J. & Jones, R. 2001. Assessment of clinical competence. *The Lancet,* 357(9260)**:**945-949.

Wayne, D.B., Fudala, M.J., Butter, J., Siddall, V.J., Feinglass, J., Wade, L.D. & McGaghie, W.C. 2005. Comparison of two standard-setting methods for advanced cardiac life support training. *Academic Medicine,* 80(10)**:**S63-S66.

Weisstein, E.W. 2014. *Bonferroni Correction*. From: MathWorld - A Wolfram web resource. Web Page: http://mathworld.wolfram.com/BonferroniCorrection.html Accessed: 5 September 2014.

WHO (World Health Organization). 2006. *Working together for health: The world health report 2006.* World Health Organization. Geneva.

WHO (World Health Organization). 2013. *Transforming and scaling up health professionals' education and training guidelines.* World Health Organization. Geneva.

Wijnen, W.H.F.W. 1971. *Onder of boven de maat (To be or not to be up to the mark).* Ph.D. thesis. Published Amsterdam: Sets & Zeitlinger.

Wilkinson, T.J., Newble, D.I. & Frampton, C.M. 2001. Standard setting in an objective structured clinical examination: use of global ratings of borderline performance to determine the passing score. *Medical Education,* 35(11)**:**1043-1049.

Wood, D.F. 2010. Formative assessment. In Swanwick, T. (Ed.) *Understanding Medical Education: Evidence, Theory, Practice* (*1st edition*). London: Wiley-Blackwell.

Wood, T.J., Humphrey-Murto, S.M. & Norman, G.R. 2006. Standard setting in a small scale OSCE: a comparison of the Modified Borderline-Group Method and the Borderline Regression Method. *Advances in Health Sciences Education. Theory and Practice.,* 11(2)**:**115-22.

Woollard, R. & Boelen, C. 2012. Seeking impact of medical schools on health: meeting the challenges of social accountability. *Medical Education,* 46(1)**:**21-27.

Wyse, A.E. 2013. Construct maps as a foundation for standard setting. *Measurement: Interdisciplinary Research and Perspectives,* 11(4)**:**139-170.

Yudkowsky, R. & Downing, S.M. 2008. Simpler Standards for Local Performance Examinations: The Yes/No Angoff and Whole-Test Ebel. *Teaching & Learning in Medicine,* 20(3)**:**212-217.

Yudkowsky, R. & Downing, S.M. 2009. Chapter 6 - Standard Setting. In Downing, S. M. & Yudkowsky, R. (Eds) *Assessment in health professions education (ebook)* (*Kindle edition*). New York: Routledge.

Zieky, M.J. 1995. *A historical perspective on setting standards*. *Proceedings of the joint conference on standard setting for large-scale assessments.* ERIC.

Zieky, M.J. 2002. Ensuring the fairness of licensing tests. *CLEAR Exam Review,* 12(1)**:**20-26.

APPENDIX A

APPENDIX A-1: ONLINE SURVEY AT TIME 1 (FEB 2013)

APPENDIX A-2: ONLINE SURVEY AT TIME 2 (FEB 2014)

APPENDIX A-1: ONLINE SURVEY AT TIME 1 (FEB 2013)

APPENDIX A-1: Online survey at Time 1 (Feb 2013)

1.

1. Dear College of Physicians of South Africa (CoP) examiner,

Many thanks indeed for taking the time to consider participation in this Ph.D. research study.
It is regarding the CoP examiners' views, attitudes, perspectives and knowledge regarding setting pass standards (or pass marks) for the written papers of the FCP(SA) Part I and II examinations.

You are invited to participate in this online survey based on your status as a CoP examiner. Please be advised that your participation in this questionnaire survey is strongly encouraged and hugely valued, but completely voluntary. The survey will be sent to all CoP examiners on the CoP database.

Ethical approval for this Ph.D. study (and this questionnaire) has been obtained from the CMSA Examinations and Credentials committee as well as the University of the Free State's Ethics committee. The supervisors of the Ph.D study are: Prof Marietjie Nel (UFS) and Prof Vanessa Burch (UCT, CMSA)

This survey is anonymous and no demographic data will be collected.

All your responses are treated confidentially and the survey will be analysed with a view to determining the examiners' views, attitudes, perspectives and knowledge about standard setting as a group. The gathered data will then be used to identify training needs and a CPSA workshop (open to all CoP examiners) will be developed to address the training needs.

Towards the end of the Ph.D. study, the views, attitudes, perspectives and knowledge of the CoP examiners will be re-evaluated (round two of the survey) to see if any significant changes have occurred as a result of the workshop and the potential exposure to the standard setting processes in the CoP.

Please note that by completing this questionnaire you are voluntarily agreeing to participate in this research study. You will remain anonymous and your data will be treated confidentially at all times. You may withdraw from this study at any given moment during the completion of the questionnaire. The researcher intends to publish the findings from the group's responses as part of his Ph.D. study. If you have any queries, please do not hesitate to contact me.

APPENDIX A-1: Online survey at Time 1 (Feb 2013)

Ph.D. researcher: Dr Scarpa Schoeman (UFS Dept of Medicine) - Email: schoemanfhs@ufs.ac.za or Cell: 0823787333

Please tick the box below to proceed to the questionnaire, IF you consent to:

· Participate in the study; and

· If your responses can be included in the data analysis

- ◯ I agree and consent to participate in the study
- ◯ I don't consent

2.

Many thanks indeed for participating in the survey.

There are 16 questions in the survey and a response to ALL the questions is required. A progress bar at the bottom of the page will indicate your progress.

If you wish to go back and edit some of your responses, please use the BACK button below the progress bar to navigate

It should take about 10-15 minutes to complete the survey

Many thanks again for your help!

Dr Scarpa Schoeman, Ph.D. researcher

3.

2. How would you rate your current knowledge about the concept of standard setting? (The process of setting the pass mark for assessments) (Choose one option)

- ◯ I know nothing about it
- ◯ I know very little about it
- ◯ I am familiar with it, but not knowledgeable about it
- ◯ I am knowledgeable about it – I can explain it to a colleague and know a few methods
- ◯ I am very knowledgeable about it and use standard setting methods in my own educational practice

Free text comments regarding this question (optional):

APPENDIX A-1: Online survey at Time 1 (Feb 2013)

3. What training have you had on the topic of standard setting? (Choose ALL the options that would describe your training most accurately)

- [ ] None
- [ ] Attended a workshop/seminar on it once
- [ ] Attended a workshop/seminar on it more than once
- [ ] Attended a programme (course) about it or where it was covered
- [ ] I have read an article about it
- [ ] I have read various articles about it
- [ ] Studied/Researched it in depth

If you attended an event about standard setting, where was it? e.g. University, CMSA, Private. Also add free text comments regarding this question (optional):

4. In 2011, the CPSA introduced the Cohen method of standard setting for the FCP(SA) Part I written MCQ examination. Are you aware of this change? (Choose one option)

- ◯ Yes
- ◯ No

Free text comments regarding this question (optional):

4.

5. In 2012, the CPSA introduced the Cohen method of standard setting for the FCP(SA) Part II Objective test examination. Are you aware of this change? (Choose one option)

- ◯ Yes
- ◯ No

Free text comments regarding this question (optional):

APPENDIX A-1: Online survey at Time 1 (Feb 2013)

6. How knowledgeable are you about the Cohen method of standard setting? (Choose one option)

- ◯ I have no idea how it works
- ◯ I have a vague idea how it works, but can't explain it to a colleague with confidence
- ◯ I know and understand it – will be able to explain or teach it to a colleague
- ◯ I would say I am an expert on it because I could / do use it myself

Free text comments regarding this question (optional):

7. The CPSA introduced the Cohen method of standard setting for the FCP(SA) Part I written MCQ examination (in 2011) and for the FCP(SA) Part II Objective test (in 2012).

Do you think it would also be appropriate for use in setting the pass mark for the two FCP(SA) Part II written SAQ (short answer question) papers? (Choose one option)

- ◯ I don't know, since I don't know the method well enough to make an informed judgement
- ◯ No - I know how the Cohen method works, but I don't think it would be suited
- ◯ No - I know how the Cohen method works, but I would rather continue using the current 50% pass mark - not keen on any changes
- ◯ No – I don't know how the Cohen method works, but I would rather continue using the current 50% pass mark - not keen on any changes
- ◯ Yes - I don't know how the Cohen method works, but I feel we need to review the current practice of a 50% pass mark
- ◯ Yes - I know how the Cohen method works and it is worth exploring its possibilities.

Free text comments regarding this question (optional):

5.

8. Prior to 2011, the pass mark for Part I written paper in the FCP(SA) examinations was fixed at 50%. What is your view on this past practice? (Choose one option)

- ◯ No problem with it
- ◯ Slightly concerned about it
- ◯ Very concerned about it
- ◯ No longer acceptable in modern educational practice

Free text comments regarding this question (optional):

APPENDIX A-1: Online survey at Time 1 (Feb 2013)

9. Prior to 2012, the pass mark for Part II Objective test in the FCP(SA) examinations was fixed at 50%. What is your view on this past practice? (Choose one option)

- ◯ No problem with it
- ◯ Slightly concerned about it
- ◯ Very concerned about it
- ◯ No longer acceptable in modern educational practice

Free text comments regarding this question (optional):

10. What is your view about the introduction of the Cohen method to set the pass mark for the FCP(SA) Part I MCQ examination in 2011? (Choose one option)

- ◯ I am deeply concerned about it, I think we should have kept the fixed 50% pass mark
- ◯ Not sure how it really works; hence I am concerned
- ◯ Not sure how it really works, but I support the CoP council's decision to introduce it
- ◯ Glad we changed to it – I fully endorse it

Free text comments regarding this question (optional):

11. What is your view about the introduction of the Cohen method to set the pass mark for the FCP(SA) Part II Objective test examination in 2012? (Choose one option)

- ◯ I am deeply concerned about it, I think we should have kept the fixed 50% pass mark
- ◯ Not sure how it really works; hence I am concerned
- ◯ Not sure how it really works, but I support the CoP council's decision to introduce it
- ◯ Glad we changed to it – I fully endorse it

Free text comments regarding this question (optional):

APPENDIX A-1: Online survey at Time 1 (Feb 2013)

12. Please rate the following statement:

I think a pass mark derived using a standard setting method, rather than a fixed 50%, should be used for all written papers in the FCP(SA) examinations (Part I, Part II written papers and objective test).

	Strongly disagree	Disagree	Uncertain	Agree	Strongly agree
Please choose ONE option	◯	◯	◯	◯	◯

Free text comments regarding this question (optional):

6.

13. Please rate the following statement:

I endorse the use of a standard setting method, as opposed to a fixed 50%, to determine the pass mark for the written papers in the FCP(SA) examinations (Part I, Part II written papers and objective test).

	Strongly disagree	Disagree	Uncertain	Agree	Strongly agree
Please choose ONE option	◯	◯	◯	◯	◯

Free text comments regarding this question (optional):

14. Please rate the following statement:

I feel neutral about using a fixed 50% pass mark for every FCP(SA) written paper (Part I, Part II written papers and objective test). That's the way we have always done it and we can continue to do so.

	Strongly disagree	Disagree	Uncertain	Agree	Strongly agree
Please choose ONE option	◯	◯	◯	◯	◯

Free text comments regarding this question (optional):

APPENDIX A-1: Online survey at Time 1 (Feb 2013)

15. Please rate the following statement:

I would be unhappy about changing the status quo of a fixed 50% pass mark for the FCP(SA) Part II written examinations (2x Short Answer Question (SAQ) papers)

	Strongly disagree	Disagree	Uncertain	Agree	Strongly agree
Please choose ONE option	◯	◯	◯	◯	◯

Free text comments regarding this question (optional):

16. Please rate the following statement:

I would support the move, if proposed, to also introduce standard setting to determine the pass mark for the FCP(SA) Part II written examinations (the 2x SAQ papers).

	Strongly disagree	Disagree	Uncertain	Agree	Strongly agree
Please choose ONE option	◯	◯	◯	◯	◯

Free text comments regarding this question (optional):

7. THANK YOU

Many thanks indeed for completing the survey!

Please click on the DONE button below to submit your responses and exit the survey.

If you wish to go back and edit some of your responses, please use the BACK button below the progress bar to navigate

If you have any questions about this Ph.D. research project, please feel free to contact me.

Dr Scarpa Schoeman, Ph.D. researcher
Email: schoemanfhs@ufs.ac.za
Cell: 0823787333

APPENDIX A-2: ONLINE SURVEY AT TIME 2 (FEB 2014)

APPENDIX A-2: Online survey at Time 2 (Feb 2014)

1.

1. Dear College of Physicians of South Africa (CPSA) examiner/moderator,

Many thanks indeed for taking the time to consider participation in this Ph.D. research study.
It is regarding the CPSA examiners/moderators' views, attitudes, perspectives and knowledge regarding setting pass standards (or pass marks) for the written papers of the FCP(SA) Part I and II examinations. This is the post-hoc and final survey 1 year after the first round.

You are invited to participate in this online survey based on your status as a CPSA examiner/moderator. Please be advised that your participation in this questionnaire survey is strongly encouraged and hugely valued, but completely voluntary. The survey will be sent to all 2010- 2012 CPSA examiners/moderators on the CPSA database.

Ethical approval for this Ph.D. study (and this questionnaire) has been obtained from the CMSA Examinations and Credentials committee as well as the University of the Free State's Ethics committee. The supervisors of the Ph.D study are: Prof Marietjie Nel (UFS) and Prof Vanessa Burch (UCT, CMSA)

This survey is anonymous and no demographic data will be collected.

All your responses are treated confidentially and the survey will be analysed with a view to determining the CPSA's views, attitudes, perspectives and knowledge about standard setting as a group. The gathered data will then be used to evaluate how the views, attitudes, perspectives and knowledge about standard setting has developed over the past 12 months given the 2 seminars that was held on the topic at the FCP council meetings in 2013 and the exposure the examiners/moderators had to standard setting during the year.

Please note that by completing this questionnaire you are voluntarily agreeing to participate in this research study. You will remain anonymous and your data will be treated confidentially at all times. You may withdraw from this study at any given moment during the completion of the questionnaire. The researcher intends to publish the findings from the group's responses as part of his Ph.D. study. If you have any queries, please do not hesitate to contact me.

Ph.D. researcher: Dr Scarpa Schoeman (UFS Dept of Medicine) - Email: schoemanfhs@ufs.ac.za or Cell: 0823787333

APPENDIX A-2: Online survey at Time 2 (Feb 2014)

Please tick the box below to proceed to the questionnaire, IF you consent to:

· Participate in the study; and

· If your responses can be included in the data analysis

◯ I agree and consent to participate in the study

◯ I don't consent

2.

Many thanks indeed for participating in the survey.

There are 16 questions in the survey and a response to ALL the questions is required. A progress bar at the bottom of the page will indicate your progress.

If you were involved in the Angoff standard setting process at one or more of the biannual FCP council meetings, you will be asked a further 2 questions.

If you wish to go back and edit some of your responses, please use the BACK button below the progress bar to navigate

It should take about 10-15 minutes to complete the survey

Many thanks again for your help!

Dr Scarpa Schoeman, Ph.D. researcher

3.

2. How would you rate your current knowledge about the concept of standard setting? (The process of setting the pass mark for assessments) (Choose one option)

◯ I know nothing about it

◯ I know very little about it

◯ I am familiar with it, but not knowledgeable about it

◯ I am knowledgeable about it – I can explain it to a colleague and know a few methods

◯ I am very knowledgeable about it and use standard setting methods in my own educational practice

Free text comments regarding this question (optional):

APPENDIX A-2: Online survey at Time 2 (Feb 2014)

3. What training have you had on the topic of standard setting? (Choose ALL the options that would describe your training most accurately)

- [ ] None
- [ ] Attended a workshop/seminar on it once
- [ ] Attended a workshop/seminar on it more than once
- [ ] Attended a programme (course) about it or where it was covered
- [ ] I have read an article about it
- [ ] I have read various articles about it
- [ ] Studied/Researched it in depth

If you attended an event about standard setting, where was it? e.g. University, CMSA, Private. Also add free text comments regarding this question (optional):

4. In 2011, the CPSA introduced the Cohen method of standard setting for the FCP(SA) Part I written MCQ examination. Are you aware of this change? (Choose one option)

- ◯ Yes
- ◯ No

Free text comments regarding this question (optional):

4.

5. In 2012, the CPSA introduced the Cohen method of standard setting for the FCP(SA) Part II Objective test examination. Are you aware of this change? (Choose one option)

- ◯ Yes
- ◯ No

Free text comments regarding this question (optional):

APPENDIX A-2: Online survey at Time 2 (Feb 2014)

6. How knowledgeable are you about the Cohen method of standard setting? (Choose one option)

- ◯ I have no idea how it works
- ◯ I have a vague idea how it works, but can't explain it to a colleague with confidence
- ◯ I know and understand it – will be able to explain or teach it to a colleague
- ◯ I would say I am an expert on it because I could / do use it myself

Free text comments regarding this question (optional):

7. The CPSA introduced the Cohen method of standard setting for the FCP(SA) Part I written MCQ examination (in 2011) and for the FCP(SA) Part II Objective test (in 2012).

Do you think it would also be appropriate for use in setting the pass mark for the two FCP(SA) Part II written SAQ (short answer question) papers? (Choose one option)

- ◯ I don't know, since I don't know the method well enough to make an informed judgement
- ◯ No - I know how the Cohen method works, but I don't think it would be suited
- ◯ No - I know how the Cohen method works, but I would rather continue using the current 50% pass mark - not keen on any changes
- ◯ No – I don't know how the Cohen method works, but I would rather continue using the current 50% pass mark - not keen on any changes
- ◯ Yes - I don't know how the Cohen method works, but I feel we need to review the current practice of a 50% pass mark
- ◯ Yes - I know how the Cohen method works and it is worth exploring its possibilities.

Free text comments regarding this question (optional):

5.

8. Prior to 2011, the pass mark for Part I written paper in the FCP(SA) examinations was fixed at 50%. What is your view on this past practice? (Choose one option)

- ◯ No problem with it
- ◯ Slightly concerned about it
- ◯ Very concerned about it
- ◯ No longer acceptable in modern educational practice

Free text comments regarding this question (optional):

APPENDIX A-2: Online survey at Time 2 (Feb 2014)

9. Prior to 2012, the pass mark for Part II Objective test in the FCP(SA) examinations was fixed at 50%. What is your view on this past practice? (Choose one option)

- ◯ No problem with it
- ◯ Slightly concerned about it
- ◯ Very concerned about it
- ◯ No longer acceptable in modern educational practice

Free text comments regarding this question (optional):

10. What is your view about the introduction of the Cohen method to set the pass mark for the FCP(SA) Part I MCQ examination in 2011? (Choose one option)

- ◯ I am deeply concerned about it, I think we should have kept the fixed 50% pass mark
- ◯ Not sure how it really works; hence I am concerned
- ◯ Not sure how it really works, but I support the CoP council's decision to introduce it
- ◯ Glad we changed to it – I fully endorse it

Free text comments regarding this question (optional):

11. What is your view about the introduction of the Cohen method to set the pass mark for the FCP(SA) Part II Objective test examination in 2012? (Choose one option)

- ◯ I am deeply concerned about it, I think we should have kept the fixed 50% pass mark
- ◯ Not sure how it really works; hence I am concerned
- ◯ Not sure how it really works, but I support the CoP council's decision to introduce it
- ◯ Glad we changed to it – I fully endorse it

Free text comments regarding this question (optional):

APPENDIX A-2: Online survey at Time 2 (Feb 2014)

12. Please rate the following statement:

I think a pass mark derived using a standard setting method, rather than a fixed 50%, should be used for all written papers in the FCP(SA) examinations (Part I, Part II written papers and objective test).

	Strongly disagree	Disagree	Uncertain	Agree	Strongly agree
Please choose ONE option	◯	◯	◯	◯	◯

Free text comments regarding this question (optional):

6.

13. Please rate the following statement:

I endorse the use of a standard setting method, as opposed to a fixed 50%, to determine the pass mark for the written papers in the FCP(SA) examinations (Part I, Part II written papers and objective test).

	Strongly disagree	Disagree	Uncertain	Agree	Strongly agree
Please choose ONE option	◯	◯	◯	◯	◯

Free text comments regarding this question (optional):

14. Please rate the following statement:

I feel neutral about using a fixed 50% pass mark for every FCP(SA) written paper (Part I, Part II written papers and objective test). That's the way we have always done it and we can continue to do so.

	Strongly disagree	Disagree	Uncertain	Agree	Strongly agree
Please choose ONE option	◯	◯	◯	◯	◯

Free text comments regarding this question (optional):

APPENDIX A-2: Online survey at Time 2 (Feb 2014)

15. Please rate the following statement:

I would be unhappy about changing the status quo of a fixed 50% pass mark for the FCP(SA) Part II written examinations (1x Short Answer Question (SAQ) paper and 1x Clinical case MCQ paper)

	Strongly disagree	Disagree	Uncertain	Agree	Strongly agree
Please choose ONE option	◯	◯	◯	◯	◯

Free text comments regarding this question (optional):

16. Please rate the following statement:

I would support the move, if proposed, to also introduce standard setting to determine the pass mark for the FCP(SA) Part II written examinations (1x Short Answer Question (SAQ) paper and 1x Clinical case MCQ paper).

	Strongly disagree	Disagree	Uncertain	Agree	Strongly agree
Please choose ONE option	◯	◯	◯	◯	◯

Free text comments regarding this question (optional):

7.

17. Were you involved in any of the Angoff standard setting meetings and procedures during one (or more) of the biannual FCP clinical examinations' meetings?

◯ Yes

◯ No

8.

APPENDIX A-2: Online survey at Time 2 (Feb 2014)

18. In your view, do you feel the Angoff method of setting the pass standard for the FCP written examinations is FEASIBLE and SUSTAINABLE to continue in the long run for the FCP examinations? Please choose ONE option and use the free text area provided to clarify your view if needed.

- ◯ Yes, I think it is feasible and we can sustain it with our resources
- ◯ Uncertain - please clarify your answer in free text area below
- ◯ No, I think it is not feasible and too resource intensive to sustain in the long run

Free text comments regarding this question:

9. THANK YOU

Many thanks indeed for completing the survey!

Please click on the DONE button below to submit your responses and exit the survey.

If you wish to go back and edit some of your responses, please use the BACK button below the progress bar to navigate

If you have any questions about this Ph.D. research project, please feel free to contact me.

Dr Scarpa Schoeman, Ph.D. researcher
Email: schoemanfhs@ufs.ac.za
Cell: 0823787333

APPENDIX B

APPENDIX B-1: STANDARD SETTING SEMINAR PRESENTATION

APPENDIX B-2: STANDARD SETTING SEMINAR EVALUATION FORM

STANDARD SETTING SEMINAR PRESENTATION

APPENDIX B-1 – Powerpoint presentation used at seminar

STANDARD SETTING FOR THE FCP (SA)

What matters most? Expert opinion vs Performance data

Dr Scarpa Schoeman

Seminar for FCP Examiners
May/Oct 2013

OVERVIEW

- Pre-seminar survey
- Introductory talk
- Discussion
- Post-seminar survey
- Highlights from online survey – Feb 2013

Pre-Seminar Survey

- Please complete the survey provided (**Before** page)
- Rate your current understanding and opinion

CMSA responsibility

- Mandate from HPCSA – national examination body needed for spes.
- From Jan 2011 – All new registrars – exit exams via CMSA
- In CoP – all registrars entry (Part I) and exit (Part II) exams
- HUGE responsibility to get the HIGH stakes exams (Part I & II) "right"
- Regulation by HPCSA and with appropriate accreditation
- Criteria for good assessment - Ottawa 2010 world consensus document

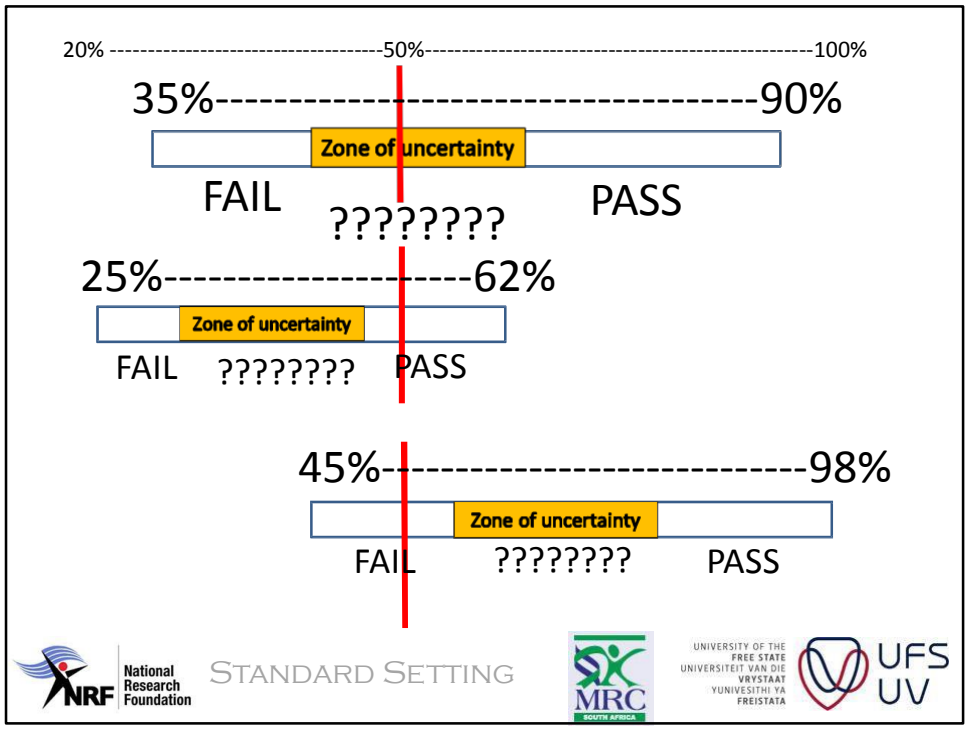
CMSA Accountability

- We need to build excellent national assessment systems
- Based on best evidence which is feasible in our context
  - Valid = aligned with outcomes
  - Reliable (reproducible) and stable
  - Fair = to candidates and patients
    - Transparency and
    - Safety minimum standard of competence
  - Credibility
  - Feasibility and
  - Favourable educational impact

CoP process to date

1. Outcomes – defined our "core" curriculum (knowledge, skills, attitudes)
2. Academic resources to cover the core (teaching & learning)
3. Assess mastery of the core
4. Standard setting system to judge mastery of core

- Do FCP exams have the same level of difficulty (standard) in each cycle?
- Assessments have varying difficulty – noise in the process (not perfect!)
- What is the pass standard??? – must logically also vary – but why?
- To make accurate decisions re the candidates
- Reduce false positives and false negatives



SUMMARY OF CONCEPT

- The "Holy Grail" in assessment!!
  - What is the *true* pass mark of the test?
  - Criteria for good assessment - Ottawa 2010 consensus
- WHY is it a critical factor in QA of assessments??
  - All tests must be calibrated – to have credible results
  - Limit **False positive** vs **False negative** outcomes
  - Critical for accurate decision making re progression
  - Implications for candidate, CoP, CMSA, HPCSA, PATIENTS

OPTIONS - METHODS

2 ways to decide on the difficulty of a written paper

- Expert judges **Angoff**
- Performance data **Cohen**

Modified Ebel 1990s

Ebel 1972

Modified Angoff 1980s

Wijnen 1971

Angoff 1971

Nedelsky 1954

Hofstee 1983

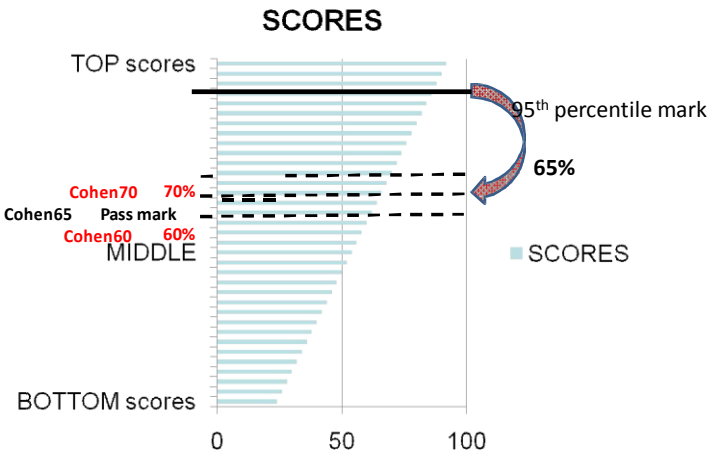
Jaeger 1983

Cohen 2010

ANGOFF METHOD

- Internationally most used and well described
- ++ modifications – due to its issues
- Expert judges must review EACH question item
- Hinges on the "Borderline student" concept
- "Would a BORDERLINE student get this item right?"
- Average the scores across items and across judges
- Feasibility challenges

THE COHEN METHOD 2010



DISCUSSION POINTS

- Variation in FCP written exams' difficulty
- Clarity of Cohen method's use and application

Post Seminar Survey

- Please complete the survey provided (**After** page)
- Rate your current understanding and opinion post seminar

ONCE EVERYBOBY IS DONE:

Next slide – presenter's opinions, given our context and resources

	Angoff	Cohen
Objectivity – how much will personal bias influence this method?	++ (based on expert opinion)	+++++ (based on expert decision)
Feasibility – is this method do-able with our expertise, time, resources?	++ (10-12 examiners, 2-4hrs each)	+++++ (1 person – 2min)
Sustainability – can we keep doing this method in the long term?	+ (costs+++)	+++++ (no cost)
Credibility – how confident are you in the results of this method?	++	+++
Validity – will this method tell us want we want to know?	+++	++++
Reliable – will this method be reproducible if repeated?	++	+++++
Transparency – how clear and understandable is this method to you and other stakeholders (public and candidates)?	++	++++

Situational Analysis of the CoP re Standard Setting

- n= 54 examiners (2010-2012)
- 38 responses = 70.4%
- Quite knowledgeable re standard setting – 84% familiar +
- 79% Part I introduced & 55% OT introduced
- 76% = all FCP exams need a SS method, and not fixed 50%
- 76% = Endorse a SS method as appose to fixed 50%
- 79% = We can't continue using a fixed 50% for FCP exams

Well done!

Situational Analysis of the CoP re Standard Setting

- Cohen Method: 84% had vague idea » expert

Exam	Question			
Part I	Fixed 50% pass mark	29%	16%	55%
	Use of Cohen method	84%	13%	3%
Part II OT	Fixed 50% pass mark	29%	18%	53%
	Use of Cohen method	87%	10%	3%
Part II SAQ	Change fixed 50% pm?	68%	13%	18%
	Support SS introduction?	74%	21%	5%
	Cohen method?	61%	26%	13%

Standard setting for the FCP (SA)

What matters most? Expert opinion vs Performance data

MANY THANKS!

Dr Scarpa Schoeman
schoemanfhs@ufs.ac.za

Seminar for FCP Examiners
May/Oct 2013

MANY THANKS!

APPENDIX B-2: STANDARD SETTING SEMINAR EVALUATION FORM

Please rate (**by circling**) the following 2 methods by the parameters on the **left column** (or √ the "**I don't know**" box)
Please provide your **opinion at this point in time**

Parameter	Angoff method	Cohen Method
Objectivity – how much will personal bias influence this method?	1 2 3 4 5 6 7 **NOT** objective **VERY** objective ☐ = I don't know	1 2 3 4 5 6 7 **NOT** objective **VERY** objective ☐ = I don't know
Feasibility – is this method do-able with our expertise, time, resources?	1 2 3 4 5 6 7 **NOT** feasible **VERY** feasible ☐ = I don't know	1 2 3 4 5 6 7 **NOT** feasible **VERY** feasible ☐ = I don't know
Sustainability – can we keep doing this method in the long term?	1 2 3 4 5 6 7 **NOT** sustainable **VERY** Sustainable ☐ = I don't know	1 2 3 4 5 6 7 **NOT** sustainable **VERY** Sustainable ☐ = I don't know
Credibility – how confident are you in the results of this method?	1 2 3 4 5 6 7 **NOT** credible **VERY** credible ☐ = I don't know	1 2 3 4 5 6 7 **NOT** credible **VERY** credible ☐ = I don't know
Validity – will this method tell us want we want to know?	1 2 3 4 5 6 7 **NOT** Valid **VERY** Valid ☐ = I don't know	1 2 3 4 5 6 7 **NOT** Valid **VERY** Valid ☐ = I don't know
Reliable – will this method be reproducible if repeated?	1 2 3 4 5 6 7 **NOT** Reliable **VERY** Reliable ☐ = I don't know	1 2 3 4 5 6 7 **NOT** Reliable **VERY** Reliable ☐ = I don't know
Transparency – how clear and understandable is this method to you and other stakeholders (public and candidates)?	1 2 3 4 5 6 7 **NOT** Transparent **VERY** Transparent ☐ = I don't know	1 2 3 4 5 6 7 **NOT** Transparent **VERY** Transparent ☐ = I don't know

APPENDIX C

INVITATION TO PARTICIPATE IN A STANDARD SETTING PANEL AND CONSENT LETTER

APPENDIX C

INVITATION TO PARTICIPATE IN A STANDARD SETTING PANEL AND CONSENT LETTER

Dear College of Physicians of South Africa (CoP) examiner

I am writing to you to request your participation in my Ph.D. study entitled: **"Standard setting for specialist physician examinations in South Africa"**

You are invited to partake in a standard setting process called the **Angoff method** to determine the pass mark for the Part I and Part II written examinations of the FCP (SA). The process of and how the Angoff method works will be explained to you by your fellow examiner and one of my promoters for this Ph.D. study, Prof Vanessa Burch. All your responses in the Angoff process will be anonymous. Furthermore, I wish to assure you that your responses in the Angoff process will be treated in a highly confidential manner. Please note that we aim to publish the findings of this research study. The Angoff process will be facilitated at each step and will take approximately 2 hours to complete.

The results from the Angoff method of setting the pass standard will be compared to that of the Cohen method. I hope that you are willing to partake in this important research project for the CoP to strengthen the quality assurance of the FCP (SA) examinations.

This Ph.D. research study is conducted through the University of the Free State (UFS) and the appropriate approval has been obtained from the Ethics Committee of the Faculty of Health Sciences, UFS (ECUFS NO.: 94/2012) and the CoP.

The promoters for my study are as follows:

Prof. M.M. Nel (Internal promoter)
Head: Division Health Sciences Education
Faculty of Health Sciences
University of the Free State

Prof. V.C. Burch (External promoter)
Professor and Chair of Clinical Medicine
Department of Medicine
Faculty of Health Sciences
University of Cape Town
Secretary of the College of Physicians of South Africa

Yours sincerely

Dr F.H. Scarpa Schoeman
Ph.D. candidate and principal investigator
Senior Lecturer/Specialist in Medical Education
Department of Internal Medicine
School of Medicine
Faculty of Health Sciences
University of the Free State
REGISTERED PROJECT
(ECUFS NO.: 94/2012)

FORM OF CONSENT TO BE COMPLETED BY ANGOFF PANELLISTS

Date: _____

Hereby I, the undersigned, consent to participate in the **Angoff standard setting process** as part of the Ph.D. study of Dr FHS Schoeman of the University of the Free State. I also undertake to ensure that my participation in this process remains confidential and that no information from the research process will be divulged. My particulars are as follows (please use BLOCK CAPITALS):

Title:……..

Surname:……

First Name:……..

Signature of consent……

All your responses in the Angoff process will be anonymous and blinded to the researcher. Furthermore, be assured you that your responses in the Angoff process will be treated in a highly confidential manner. Please note that we aim to publish the findings of this research study.

Please return this form to the convenor of the FCP (SA) examination, who will return the form to Dr Schoeman.

APPENDIX D

CODING PROCEDURE USED TO PRODUCE TABLE 4.7 AND TABLE 4.8

APPENDIX D

CODING PROCEDURE USED TO PRODUCE TABLE 4.7 AND TABLE 4.8

Ten of the survey items, 7 – 16, addressed the attitudes, views and perspectives of the CoP examiners about standard setting. The data from these items were coded and re-organised to enable effective analysis and comparison between the T1 and T2 rounds of the survey.

There were two types of items used in the survey:

1. **MCQ**-type: Pick the single most appropriate option from a multiple set of options.
2. **Likert**-type: Rate the item's statement on a 5-point Likert-rating scale.

The data from the Likert-items were subsequently clustered three categories:

1. **"Agreement"** (strongly agree + agree)
2. **"Uncertain"**
3. **"Disagreement"** (strongly disagree + disagree)

Each survey item addressed one of the three broad "**Topics of Change**" areas as a main focus. The three "Topics of Change" areas were (*cf.* Tables 4.7 and 4.8 in Chapter 4):

- Changing the previous fixed 50% pass mark
- The current use of standard setting (the Cohen method)
- The expanded use of standard setting (the Cohen method)

The individual raw data from each of these items are provided below, per "topic of change", in item order. The data components within each item were coded as:

- **S** – (Supporting change)
- **U** – (Uncertain of change)
- **R** – (Resisting change)

Changing the previous fixed 50% pass mark
3 Items

Item 8 and 9

MCQ-type items

Views about past use of fixed 50% pass mark in Part I (MCQ) and Part II (OT)

50% pass mark views	Exam	Response n (%)		*Code*
		T1	T2	
No problem with it	MCQ	11 (29%)	5 (12%)	**R**
	OT	11 (29%)	5 (12%)	**R**
Slightly concerned about it	MCQ	6 (16%)	4 (10%)	**U**
	OT	7 (18%)	4 (10%)	**U**
Very concerned about it	MCQ	6 (16%)	12% (5)	**S**
	OT	5 (13%)	12% (5)	**S**
No longer acceptable in modern educational practice	MCQ	15 (40%)	27 (66%)	**S**
	OT	15 (40%)	27 (66%)	**S**

OT – Objective Test

Item 14

Likert-type item

Views about use of fixed 50% pass mark in all FCP (SA) written exams

Please rate the following statement: I feel neutral about using a fixed 50% pass mark for every FCP (SA) written paper (Part I, Part II written papers and objective test). That's the way we have always done it and we can continue to do so.

Rating label	Ratings (n)		Code
	T1	T2	
Disagreement	**30**	**34**	**S**
Strongly disagree	9	19	
Disagree	21	15	
Uncertain	**5**	**2**	**U**
Agree	3	3	
Strongly agree	0	2	
Agreement	**3**	**5**	**R**

The current use of standard setting (the Cohen method) 4 Items

Item 10 and 11

MCQ-type items

Views about introducing the Cohen Method for the Part I (MCQ) and Part II (OT)

Cohen Method's introduction	Exam	Response n (%)		*Code*
		T1	T1	
I am deeply concerned about it, I think we should have kept the fixed 50% pass mark	MCQ	1 (3%)	3 (7%)	**R**
	OT	1 (3%)	3 (7%)	**R**
Not sure how it really works; hence I am concerned	MCQ	5 (13%)	3 (7%)	**U**
	OT	4 (11%)	3 (7%)	**U**
Not sure how it really works, but I support the CPSA council's decision to introduce it	MCQ	8 (21%)	4 (10%)	**S**
	OT	9 (24%)	3 (7%)	**S**
Glad we changed to it – I fully endorse it	MCQ	24 (63%)	31 (76%)	**S**
	OT	24 (63%)	32 (78%)	**S**

OT – Objective Test

Item 12

Likert-type item

Support for using standard setting rather than a fixed 50% pass mark

Please rate the following statement: I think a pass mark derived using a standard setting method, rather than a fixed 50%, should be used for all written papers in the FCP(SA) examinations (Part I, Part II written papers and objective test).

Rating label	Ratings (n)		Code
	T1	T2	
Disagreement	**4**	**4**	**R**
Strongly disagree	0	2	
Disagree	4	2	
Uncertain	**5**	**2**	**U**
Agree	21	16	
Strongly agree	8	19	
Agreement	**29**	**35**	**S**

Item 13

Likert-type item

Endorsement of using standard setting rather than a fixed 50% pass mark

Please rate the following statement: I endorse the use of a standard setting method, as opposed to a fixed 50%, to determine the pass mark for the written papers in the FCP (SA) examinations (Part I, Part II written papers and objective test).

Rating label	Ratings (n)		Code
	T1	T2	
Disagreement	**4**	**4**	**R**
Strongly disagree	0	1	
Disagree	4	3	
Uncertain	**5**	**2**	**U**
Agree	22	15	
Strongly agree	7	20	
Agreement	**29**	**35**	**S**

The expanded use of standard setting (the Cohen method) 3 Items

Item 7

MCQ-type items

Views on the expanded use of the Cohen Method (in the SEQ exam)

Cohen method expanded use	**Response n (%)**		***Code***
	T1	**T2**	
I don't know, since I don't know the method well enough to make an informed judgement	10 (26%)	2 (5%)	**U**
No - I know how the Cohen method works, but I don't think it would be suited	2 (5%)	4 (10%)	**R**
No - I know how the Cohen method works, but I would rather continue using the current 50% pass mark - not keen on any changes	2 (5%)	1 (2%)	**R**
No – I don't know how the Cohen method works, but I would rather continue using the current 50% pass mark - not keen on any changes	1 (3%)	2 (5%)	**R**
Yes - I don't know how the Cohen method works, but I feel we need to review the current practice of a 50% pass mark	4 (11%)	3 (7%)	**S**
Yes - I know how the Cohen method works and it is worth exploring its possibilities.	19 (50%)	29 (71%)	**S**

Item 15

Likert-type item

Resistance to change the status quo of 50% pass mark for SEQ exam

Please rate the following statement: I would be unhappy about changing the status quo of a fixed 50% pass mark for the FCP(SA) Part II written examinations (2x Short Answer Question (SEQ) papers)

Rating label	Ratings (n)		Code
	T1	T2	
Disagreement	**26**	**34**	**R**
Strongly disagree	8	15	
Disagree	18	19	
Uncertain	**5**	**1**	**U**
Agree	7	3	
Strongly agree	0	3	
Agreement	**7**	**6**	**S**

Item 16

Likert-type item

Support for introduction of standard setting for SAQ exam

Please rate the following statement: I would support the move, if proposed, to also introduce standard setting to determine the pass mark for the FCP(SA) Part II written examinations (the 2x SEQ papers).

Rating label	Ratings (n)		Code
	T1	T2	
Disagreement	**2**	**4**	**R**
Strongly disagree	0	1	
Disagree	2	3	
Uncertain	**8**	**0**	**U**
Agree	21	21	
Strongly agree	7	16	
Agreement	**28**	**37**	**S**

APPENDIX E

ITEM QUALITY INDEX PLOT EXPLANATION

APPENDIX E

ITEM QUALITY INDEX PLOT EXPLANATION

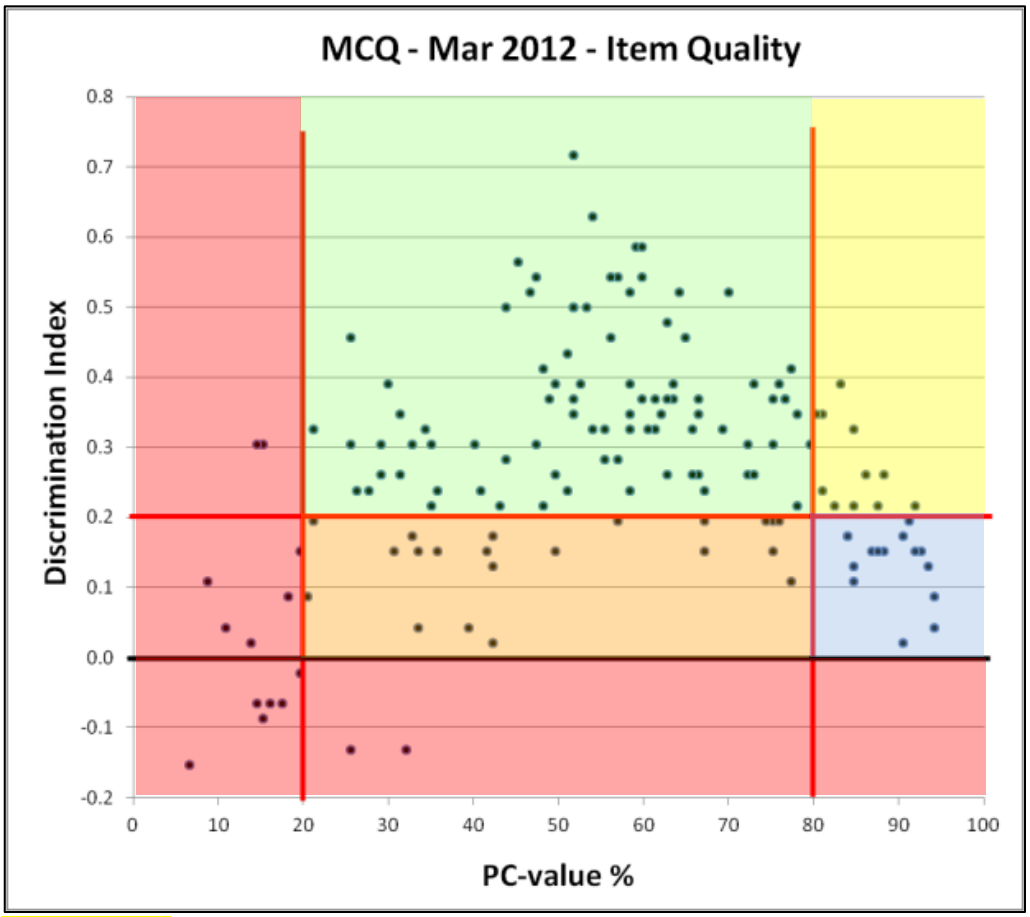


FIGURE 5.2(a): MCQ - MARCH 2012

Green zone

RETAIN in bank - Good (acceptable) quality items

Orange zone

These items have good potential, but need **REPAIR** to improve discriminant ability.

Yellow zone

REVIEW and decide what proportion is appropriate for level and training of examinee cohort

Blue zone

REMOVE – too easy, not discriminant (if topic important, rewrite item to improve properties)

Red Zone

Problematic items - Discard or **REWRITE** completely if topic is important

APPENDIX F

ETHICAL APPROVAL AND PERMISSION TO CONDUCT THE STUDY

UNIVERSITY OF THE FREE STATE
UNIVERSITEIT VAN DIE VRYSTAAT
YUNIVESITHI YA FREISTATA
UFS UV

Research Division
Internal Post Box G40
☎(051) 4052812
Fax (051) 4444359

E-mail address: StraussHS@ufs.ac.za

Ms H Strauss/hv

2012-07-30

REC Reference nr 230408-011
IRB nr 00006240

DR FHS SCHOEMAN
DEPT OF INTERNAL MEDICINE
FACULTY OF HEALTH SCIENCES
UFS

Dear Dr Schoeman

ECUFS NR 94/2012
DR FHS SCHOEMAN **DEPT OF INTERNAL MEDICINE**
PROJECT TITLE: STANDARD SETTING FOR SPECIALIST PHYSICIAN EXAMINATIONS IN SOUTH AFRICA.

- You are hereby kindly informed that at the meeting on 24 July 2012 the Ethics Committee approved the above study after the following were submitted.
  - ***The signed permission letters from the authorities***
- Committee guidance documents: Declaration of Helsinki, ICH, GCP and MRC Guidelines on Bio Medical Research. Clinical Trial Guidelines 2000 Department of Health RSA; Ethics in Health Research: Principles Structure and Processes Department of Health RSA 2004; Guidelines for Good Practice in the Conduct of Clinical Trials with Human Participants in South Africa, Second Edition (2006); the Constitution of the Ethics Committee of the Faculty of Health Sciences and the Guidelines of the SA Medicines Control Council as well as Laws and Regulations with regard to the Control of Medicines.
- Any amendment, extension or other modifications to the protocol must be submitted to the Ethics Committee for approval.
- The Committee must be informed of any serious adverse event and/or termination of the study.
- A progress report should be submitted within one year of approval of long term studies and a final report at completion of both short term and long term studies.
- Kindly refer to the ECUFS reference number in correspondence to the Ethics Committee secretariat.

Yours faithfully

.....
for CHAIR: ETHICS COMMITTEE

REQUEST FOR PERMISSION TO CONDUCT Ph.D. STUDY

Dear **Prof Gert van Zyl - Dean, UFS Faculty of Health Sciences**

With this letter I hereby formally request permission to conduct a Ph.D. research study in collaboration with the College of Physicians of South Africa (CPSA), which is one of the largest colleges within the Colleges of Medicine of South Africa (CMSA) organisation.

The title of the Ph.D. study is:
Standard setting for specialist physician examinations in South Africa

Overall goal

The overall goal of the study is to improve the quality (explicability, defensibility, stability, and acceptability) of the examinations of the CPSA and show the way forward for setting standards for other training programmes in health sciences in South Africa.

Aim of the study

The aim of this study is to evaluate standard setting for specialist physician examinations in South Africa. This will be done by:

- Investigating the attitudes, knowledge and perspectives of examiners in the CPSA regarding standard setting, providing appropriate training and evaluating the impact of the training on their attitudes and perspectives.

- Evaluating the explicability, defensibility, stability, and acceptability of the Cohen method to determine the pass mark of the written examinations of the CPSA.

- Comparing the outcome of the Cohen Method to the more established Angoff method.

The promoters for my study are as follows:

Prof. M.M. Nel (Internal promoter)
Head: Division Health Sciences Education
Faculty of Health Sciences
University of the Free State

Prof. V.C. Burch (External promoter)
Professor and Chair of Clinical Medicine
Department of Medicine
Faculty of Health Sciences
University of Cape Town
Secretary of the College of Physicians of South Africa

I look forward to hearing back from the committee.

Yours sincerely

Dr F.H. Scarpa Schoeman
Ph.D. candidate and principal investigator
Senior Lecturer/Specialist in Medical Education
Department of Internal Medicine (G73)
School of Medicine
Faculty of Health Sciences
University of the Free State
REGISTERED PROJECT
(ECUFS NO.: 94\2012)

Approved
Dean
12/6/12

Heg asseblief die protokol vir die studie hierby aan, asook die Etiekkomitee aansoekvorm.

Neem asb kennis dat dit die verantwoordelikheid van die navorser(s) is om te verseker dat alle toepaslike handtekeninge verkry word voor hierdie getekende vorm terugbesorg word aan die Etiekkomitee Administratiewe kantoor (D115) Francois Retief-gebou, Fakulteit Gesondheidswetenskappe, UV. Die protokol mag intussen ingehandig word vir Etiekkomitee goedkeuring terwyl handtekeninge bekom word.

A.

Approved / Goedgekeur	Rejected / Afgekeur

HEAD OF SCHOOL / HOOF VAN DIE SKOOL

SIGNATURE / HANDTEKENING **DATE / DATUM**

COMMENTS / KOMMENTAAR:

B.

Approved / Goedgekeur	Rejected / Afgekeur

Reeds getekend toestemmings brief ontvang

DEAN OF THE FACULTY / DEKAAN VAN DIE FAKULTEIT

SIGNATURE / HANDTEKENING **DATE / DATUM**

COMMENTS / KOMMENTAAR:

C.

Approved / Goedgekeur	Rejected / Afgekeur

Awaiting

PROF. HR HAY
VICE-RECTOR: ACADEMIC
TEL: 051-4013773

VICE-RECTOR: ACADEMIC
VISE-REKTOR: AKADEMIES /

SIGNARURE / HANDTEKENING **DATE / DATUM/**

COMMENTS / KOMMENTAAR:

D.

Approved / Goedgekeur	Rejected / Afgekeur

(If research will include students on campus and if questionnaires will be distributed in hostels on campus)

N/A

DEAN: STUDENT AFFAIRS /
DEKAAN:
STUDENTE AANGELEENTHEDE

SIGNARURE / HANDTEKENING **DATE / DATUM/**

COMMENTS / KOMMENTAAR:

E. ***INSTITUTIONAL APPROVAL*** (Institutional approval by both the Research and Ethics and Biosafety Committees must accompany each application.)

Research Committee

Ethics and Biosafety Committees

RESEARCH COMMITTEE (Please complete for each application)

Name	Dr Scarpa Schoeman
Project	Standard Setting for Specialist Physician examinations in South Africa
Institutional Support	The University of the Free State, Faculty of Health Sciences' Research committee fully supports this PhD research study of Dr Schoeman. Our internal Faculty PhD evaluation committee and Faculty board approved the project.

.....
Signature of Representative of Institution

12/9/2012
.....
Date

ETHICS AND BIOSAFETY COMMITTEES (Please complete for each application)

Name	Dr Scarpa Schoeman
Project	Standard Setting for Specialist Physician examinations in South Africa
Reviewed by Ethics Committee	This PhD research study by Dr Schoeman was formally reviewed and approved by the Ethics Committee of the UFS, Faculty of Health Sciences. The ethics committee number for the study in **94/2012**. See approval letter attached.
Reviewed by Biosafety Committee	**NA**

.....
Signature of Chairman of Ethics Committee

PROF. W.H. KRUGER
2012 -09- 12
.....
Date

CMSA

Incorporated Association not for gain (Reg. No. 1955/000003/08)
Nonprofit Organisation (Reg No 009-874 NPO)
27 Rhodes Ave, PARKTOWN WEST 2193
Private Bag X23, BRAAMFONTEIN 2017
Tel: +27 11 726-7037/8/9
Fax: +27 11 726-4036
General: admin@cmsa-jhb.co.za
Academic Registrar: alv@cmsa-jhb.co.za
Website: http://www.collegemedsa.ac.za

21 August 2012

Ref: Fac 1 B

Prof V Burch
J Floor
Department of Medicine
Old Main Building
Groote Schuur Hospital
Main Road
OBSERVATORY
7925

Dear Professor Burch

REQUEST TO CONDUCT A PHD STUDY IN THE COLLEGE OF PHYSICIANS
Your letter of 9 August 2012 refers

It gives me great pleasure to inform you that the Examinations and Credentials committee at its recent meeting approved the proposal for a study to be conducted in the College of Physicians.

Please let me know at your earliest convenience the raw data that you will need for this study.

Yours sincerely

Mrs Ann Vorster
ACADEMIC REGISTRAR

ALV/ab