ANALYSIS OF PRE-AND-POST-DISCUSSION ANGOFF RATINGS FOR EVIDENCE OF SOCIAL INFLUENCE EFFECTS

Catherine Wessen
B.A., California State University, Sacramento, 2007

THESIS

Submitted in partial satisfaction of the requirements for the degree of

MASTER OF ARTS

in

PSYCHOLOGY
(Industrial/Organizational Psychology)

at

CALIFORNIA STATE UNIVERSITY, SACRAMENTO

SUMMER

2010
ANALYSIS OF PRE-AND-POST-DISCUSSION ANGOFF RATINGS FOR EVIDENCE OF SOCIAL INFLUENCE EFFECTS

A Thesis

by

Catherine Wessen

Approved by:

__________________________, Committee Chair
Gregory Hurtz, Ph.D.

__________________________, Second Reader
Lisa Harrison, Ph.D.

__________________________, Third Reader
Lisa M. Bohon, Ph.D.

Date: ____________________________
Student: Catherine Wessen

I certify that this student has met the requirements for format contained in the University format manual, and that this thesis is suitable for shelving in the Library and credit is to be awarded for the thesis.

__________________________, Graduate Coordinator
Jianjian Qin, Ph.D

Date

Department of Psychology
Abstract

of

ANALYSIS OF PRE-AND-POST-DISCUSSION ANGOFF RATINGS FOR EVIDENCE OF SOCIAL INFLUENCE EFFECTS

by

Catherine Wessen

The Angoff (1971) method is a widely used standard-setting approach in test development. It relies on Subject Matter Experts (SMEs) who examine the difficulty level of test items and then give probability estimates on item difficulty (Hurtz & Hertz, 1999). However, expert judgment is subject to social influences that may compromise the accuracy of the desired passing score. Archival data was utilized to investigate how social influence effects might have guided the judgments of SMEs (N = 260) in setting cutoff points for 38 licensure examinations. The results showed significant changes between SME’s’ pre-and-post-discussions standard deviations, indicating that SMEs’ changed their final ratings inward toward the mean of the group. There was no support for group polarization effects. There was support for the minority pressure influence. There was also support for the insufficient anchor and adjustment bias. These results indicated that social influence factors are present in Angoff workshops.

________________________, Committee Chair
Gregory Hurtz, Ph.D.
ACKNOWLEDGMENTS

I would like to give my sincere gratitude to Dr. Gregory Hurtz who worked with me tirelessly and patiently throughout this entire process. His help and constructive input were invaluable in the completion of this thesis. I would also like to thank Dr. Lisa Harrison, and Dr. Lisa Bohon who at times believed in me more than I believed in myself. I will be forever grateful for the knowledge that I have gained from all of you at CSUS. I also like to give my sincere gratitude to the Department of Consumer Affairs - Office of Professional Examination Services for giving me access to and letting me use their data for my study. I would like to give special thanks to my supervisor Amy Welch, and my friends and coworkers Maria Avalos, Devika Tandan, Irene Wong-Chi, and George Yorba for going out of their way to help me as I juggled family, school, and work. Above all, I would like to give my heartfelt thanks to my husband Paul, and my children Wesley and Nathan for their sincere support and patience throughout my educational journey. Thanks for being there when I needed you.
# TABLE OF CONTENTS

<table>
<thead>
<tr>
<th>Chapter</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Acknowledgments</td>
<td>vi</td>
</tr>
<tr>
<td>List of Tables</td>
<td>ix</td>
</tr>
<tr>
<td>List of Figures</td>
<td>x</td>
</tr>
<tr>
<td><strong>Chapter</strong></td>
<td></td>
</tr>
<tr>
<td>1. INTRODUCTION</td>
<td>1</td>
</tr>
<tr>
<td>2. CRITERION-REFERENCED TESTING</td>
<td>3</td>
</tr>
<tr>
<td>3. SETTING PERFORMANCE STANDARDS ON CRITERION-REFERENCED TESTS</td>
<td>8</td>
</tr>
<tr>
<td></td>
<td>Person-Centered Methods</td>
</tr>
<tr>
<td></td>
<td>Item-Centered Methods</td>
</tr>
<tr>
<td>4. APPLICATIONS OF SOCIAL PSYCHOLOGY TO THE STANDARD SETTING PROCESS</td>
<td>23</td>
</tr>
<tr>
<td></td>
<td>Cognitive Biases that can Interfere with SMEs Judgment</td>
</tr>
<tr>
<td></td>
<td>Social Influences that can Interfere with SMEs Judgment</td>
</tr>
<tr>
<td>5. The CURRENT STUDY</td>
<td>48</td>
</tr>
<tr>
<td></td>
<td>Observable Patterns of Conformity Pressure, Group Polarization, Minority Pressure, and the Insufficient Anchor and Adjustment Influences</td>
</tr>
<tr>
<td>6. METHOD</td>
<td>54</td>
</tr>
<tr>
<td></td>
<td>Research Setting</td>
</tr>
<tr>
<td></td>
<td>Test Content</td>
</tr>
<tr>
<td></td>
<td>Judges</td>
</tr>
<tr>
<td></td>
<td>Standard Setting Procedure</td>
</tr>
<tr>
<td></td>
<td>Data</td>
</tr>
<tr>
<td></td>
<td>Statistical Analyses</td>
</tr>
<tr>
<td></td>
<td>Results</td>
</tr>
</tbody>
</table>
Addressing the Non Independence Assumption .................................................. 75
General Discussion ................................................................................................ 76
References ............................................................................................................. 84
LIST OF TABLES

1. Table 1 Descriptive Statistic for Initial Ratings, Final Ratings, and Item $p$-value

2. Table 2 Analysis of Variance for the Degree to which Pre-to-post-discussion Change Varies across Examinations

3. Table 3 Analysis of Variance for the Degree to which Changes in Standard Deviation Varies across Examination from Pre-to-post-discussion
**LIST OF FIGURES**

<table>
<thead>
<tr>
<th>Figure</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.</td>
<td>Figure 1 Mean Changes across Examination Cutoff Scores</td>
<td>66</td>
</tr>
<tr>
<td>2.</td>
<td>Figure 2 Standard Deviation Changes from Initial (1) to Final (2) Ratings</td>
<td>69</td>
</tr>
<tr>
<td></td>
<td>across Examination Cutoff Scores</td>
<td></td>
</tr>
<tr>
<td>3.</td>
<td>Figure 3 Correlation between Initial Ratings and Change Score</td>
<td>71</td>
</tr>
<tr>
<td>4.</td>
<td>Figure 4 Differences in Ratings from pre-to-post-Discussion</td>
<td>72</td>
</tr>
</tbody>
</table>
Since criterion-referenced scores require making vital decisions that affect the lives of people, it is critical that examination developers exercise good judgment in making such crucial decisions. Despite continuous efforts to improve standard setting methodology in testing, methods of establishing cutoff scores are based on subjective judgments and have been frequently criticized as being prone to error due to cognitive biases and social influences. Criticisms are also directed at what constitutes an appropriate balance between passing those who should fail and failing those who should pass (Zieky, 2001).

The Angoff Method, which is the focus of this study, is a widely used standard-setting approach in test development. It is used in a variety of settings such as education, licensure, and employment selection. The Angoff method is rooted in the theory that a level of minimum competency can be determined by a panel of judges, who are Subject Matter Experts (SMEs), in setting a cutoff score. Angoff (1971) contended that this minimum level of competency can be established based on the knowledge, skills, and abilities (KSAs) necessary that a randomly selected test taker would be able to get the test question correct.

The Angoff method relies on Subject Matter Experts (SME) who examine the content of each test question (item) and then estimated the probability that a minimally competent candidate (MCC) would be able to answer the item correctly (Hurtz & Hertz, 1999). If there is a discrepancy in the initial ratings of a test item, SMEs are often given a
second chance to discuss and then modify their ratings. The individual rating estimates of each test item are summed and then divided by the number of SMEs. The result is the passing score (Ferdous & Plake, 2007).

This study investigated how cognitive biases and social influence factors might influence the decisions of SMEs in an Angoff passing score workshop. Archival data of several Angoff passing score workshops were analyzed for evidence of social influence effects. More specifically, the study looked at changing patterns between the initial and final ratings of SMEs.

In this study, the term “passing score” assumes the same meaning as “cutoff score”. This is a point in the predictor distribution below which test takers are rejected and above which they are considered as successful and competent to be licensed to practice.

Chapter 1 gives an overview of this study. Chapter 2 discusses what criterion-referenced testing is; how it is used in testing, how it is distinguished from norm-referenced testing; and how such testing is developed. Chapter 3 discusses the different ways of setting performance standards for criterion-referenced tests. It also discusses the strengths and weakness of these alternative methods. Chapter 4 reviews the different ways that social cognition and group processes can influence standard setting processes and outcomes. Chapter 5 specifies the scope of the current study; including its importance, hypotheses, independent variables, and dependent variables. Chapter 6 explains the methodology, types of analyses that were used in the study, results of the study and general discussion.
According to Poggio, Glasnapp, and Eros (1981), the concept of criterion-referenced testing was first introduced in the 1960s and since then, criterion-referenced measures have been the primary focus of attention among examination developers. In the early 1970s, the widespread adoption of criterion-referenced measures was reflected in minimum competency testing (Poggio et al., 1981). The purpose of minimum competency testing is to differentiate between individuals who are knowledgeable in a specific content area versus those who are not (McGinty, 1996). Several criterion-referenced measures are used in minimum competency testing in order to determine whether an examinee meets a certain specific performance goal. The purpose of criterion-referenced testing (CRT) is to measure the depth of knowledge and skills acquired by an individual in a specific curriculum (Shepard, 1979). Criterion-referenced testing specifically measure skills that make up a particular curriculum which can be recognized by experts in that specific field of study. With criterion-referenced testing, each test taker is compared with a preset standard for acceptable achievement and the performance of other test takers is not considered or taken into consideration (Popham, 1975).

In contrast, norm-referenced testing (NRT) is a type of measure that is utilized in measuring several skill areas, for example that are sampled from a variety of textbooks and syllabi (Shepard, 1979). Bond (1996) explains that the primary purpose of norm-referenced testing is to classify test takers. Norm-referenced testing measures are designed to draw attention to achievement differences between and among test takers so
as to produce a dependable rank order of test takers across a continuum of achievement from high achievers to low achievers (Bond, 1996). Norm-referenced testing is used to compare and rank test takers with respect to the achievement of others in broad areas of knowledge (Popham, 1975).

In norm-referenced testing, a representative group of test takers, or norm group, is given a test prior to its availability to the public. The scores of the examinees who take the test after publication are then compared to those of the norm group (Bond, 1996). Because norming a test can be an expensive and complex process, test publishers typically use norms for seven-year periods, during which time the scores of all individual test takers are compared to the original norm group (Bond, 1996). Several tests such as the California Achievement Test (CTB/McGraw-Hill) and the Metropolitan Achievement Test (Psychological Corporation) are normed using a national sample of test takers (Bond, 1996). In norm-referenced testing, the performance of each test taker is compared to that of others, and a score that is usually expressed as a grade or percentile is then assigned (Popham, 1975).

Testing is an important tool that provides valuable information that enables test developers to assess the necessary knowledge required in a variety of settings (Poggio, et al. 1981). Some of the settings in which testing is prevalent include education, employment, licensure, and clinical just to name a few. Many professional examination developers apply criterion-referenced cutoff scores as standards for competency practice to all candidates regardless of the form of the administration of the examination.
According to the Office of Professional Examination Services (OPES), the adoption of a criterion-referenced passing score may eventually result in the licensure of candidates who have adequate knowledge and experience to guarantee public health and safety (OPES, Informational series, 4, 2003). Similarly, Hurtz and Auerbach (2003) assert that setting cutoff score is important because classifying people in the categories of pass or fail establishes whether an individual is considered competent to obtain a license or certification to practice in an occupation, or if the individual is knowledgeable enough to acquire employment in a particular job.

Zeiky (2001) contends that there are specific situations in which there is no other choice but to establish cut-scores. An example of such a situation is the competency tests used for licensing professional examinations in which cut-score setting is mandated by law (Zieky, 2001). Because criterion-referenced scores are used in making vital decisions that affect the lives of people, it is critical that policymakers utilize good judgment in making such important decisions. As Zieky and Perie (2006) suggested, policymakers need to consider both the negative and positive effects that cutoff scores may have on examinees. They also suggest that policymakers should investigate how examination decisions are presently made and whether determining a cutoff score would result in improved decision-making (Zieky & Perie, 2006).

There are several processes or steps involved in the development of criterion-referenced testing. In professional or licensure examinations, the first step usually begins with an occupational analysis or job analysis of the chosen occupation. An occupational analysis is a comprehensive description and evaluation of the current practice in a job.
The process of an occupational analysis begins with the selection and interviewing of a diverse sample of practitioners who accurately represent the spectrum of practices or specialty mix of the chosen profession. Practitioners are asked to identify the tasks that they perform within major categories of their profession and the knowledge they need to perform a specific task. A questionnaire based on statements of job activities is then sent to a representative sample of licensed practitioners. Once the questionnaires are completed and returned, the data are analyzed and the results are used to update the description of the practice and/or develop an examination outline (OPES, Informational Series 1, 2003).

The second step involved in the development of criterion-referenced testing is to use the examination outline that has been created as a result of the occupational analysis to write test questions for the chosen profession. The occupational analysis establishes the job-relatedness of the examination and thereby provides empirical linkage to the content of the examination. The examination outline serves as the blueprint for the development of examination questions. Test questions are constructed by subject matter experts and are critically reviewed by a different group of subject matter experts in the field to ensure that the questions are directly relevant to tasks performed on the job. After this review, the test can be administered to candidates (OPES, Informational Series 1, 2003).

The next step in the development of criterion-referenced testing in licensure examination is setting standards for the chosen profession. The setting of examination cutoff scores is done by a group of subject matter experts who are knowledgeable about
the population of the candidates and the test material (Mills, Melican, & Ahluwalia, 1991). The cutoff scores are used to classify test takers into pass or fail performance levels. Once the examination is administered, an item analysis is conducted to evaluate the performance of the newly created test questions. The development of criterion-referenced testing in licensure examination is typically conducted every five to seven years.
Chapter 3

SETTING PERFORMANCE STANDARDS ON CRITERION-REFERENCED TESTS

The setting of criterion-referenced standards is a process that usually involves practitioners who are representative of all aspects of a practice or profession. The process of establishing cutoff scores requires a panel of judges or subject matter experts who are knowledgeable about the material, the candidate population, and the methodology that is going to be used (Cizek & Bunch, 2007; Mills, Melican, & Ahluwalia, 1991). To ensure that the description of the profession represents the job tasks of professionals entering the profession, the standard setting process often includes a number of newly licensed practitioners (OPES, Informational Series, 4, 2003).

An important issue that arose within competency testing (i.e., ability or capability testing) methods was that of determining the performance standards or passing score (Poggio et al. 1981). The first few steps of any standard setting procedure include an introduction and orientation to the assessment instrument and to the standard setting method. Typically, a training session is carried out in which the panelists are educated about the standard setting method, a commonly accepted definition of minimal competence, and alternative ways to rate test items (Mills et al. 1991). Criterion-referenced standard setting begins with a conceptual definition of what a minimally acceptable level of competence for safe practice that candidates must possess in order to perform competently and safely in their work (OPES, Informational Series, 4, 2003).
Next, the panel develops common definitions of different levels of candidate performance by identifying critical work behaviors that contrast the highly effective, the minimally competent, and the ineffective candidate (OPES, Informational Series, 4, 2003). These differences are typically defined according to observable work behaviors through which candidates can display their knowledge, skills, or abilities. The raters utilize an occupational analysis as the basis for the development of behavioral descriptions of each competency level, and must then reach a consensus on meanings and wordings (OPES, Informational Series, 1, 2003). Next, a training session is presented on how to assign ratings for the test items and then raters are asked to evaluate a subset of test items in a calibration session.

After minimum competency standards and test items have been developed, a cutoff score must be determined. A variety of methods has been proposed to determine cutoff scores over the last 40 years. Among these are the Nedelsky, Ebel, Angoff, Bookmark, and the Jaeger methods. According to Stephenson and colleagues (2000), these different methods all require judges, panelists or subject matter experts (SMEs) to make a decision on how a test taker who meets the minimum qualification of the job or profession would answer the items on the test correctly. These methods can also be applied before or after the administration of the test (Stephenson, Elmore, & Evans (2000). Although all these different cutoff methods have been utilized, the Angoff (1971) method has been researched widely and is the most common standard setting procedure used by most practitioners (Ferdous & Plake, 2007).
Because several important decisions are based on test results, it is imperative that certain standards are set for testing. Standard setting can be defined in several contexts. For example, Cizek (1993) describes standard setting in the field of education and personnel assessment as the task of achieving levels of performance in which decision or classification of examinees are made.

Furthermore, Cizek (2001) asserted that a standard-setting method must be an organized system that collects judgments of qualified individuals about the level of knowledge and skills needed for someone to be classified as above or below a standard. To do this, it is crucial that individuals understand how performance standards or passing scores are set. It is also imperative that examination developers have a specific definition of what performance standards are. It is particularly important to distinguish the standard-setting methodology from the standard itself. Typically, a standard is considered the result while standard-setting methodology is the means for achieving that result.

In licensure or certification testing arena, Cohen, Kane, Crooks, (1999) describe standard setting as an unambiguous decision that assigns test takers to one of many categories of performance based on their test score. The standard-setting method is important because it establishes the point on a test score measurement that determines the pass or fail decision based on each test taker’s examination score (Haladyna & Hess, 1999).

Because of the need to address examination-related issues, the Standards for Educational and Psychological Testing (1999) was published. The Standards provide detailed procedures for creating valid and reliable tests. It also looks at forms of validity
as well as evidence that can be used to support claims of validity. The *Standards* explain the importance of having a cutoff score for a professional licensure examination and it states that, “the cut score represents an informed judgment that those scoring below it are likely to make serious errors for want of the knowledge or skills tested” (p.53).

Furthermore, the *Standards* also make available important guidance on standard setting. For example, the AERA, APA, and NCME (1999) *Standard 4.21* states that:

> When cut scores defining pass-fail or proficiency categories are based on direct judgments about the adequacy of item or test performances or performance levels, the judgmental process should be designed so that judges can bring their knowledge and experience to bear in a reasonable way. (p. 60).

Stephenson, Elmore, and Evans (2000) contend that because standard setting methods are based on the idea that upper level test takers are categorized as high scorers and those in the lower level as low scorers, the cut score should be one that a person of borderline or minimal skills is expected to get. The *Standards*, however, warned that there is not a particular preferred method used in determining cut scores, and that there can not be a specific set of procedures for establishing their defensibility (*Standards for Educational and Psychological Testing, 1999*).

The Office of Professional Examination Services (OPES) emphasizes that although the purpose of academic examinations is to evaluate how well a person can define and comprehend terms and concepts, the one and only purpose of a licensing examination is to identify persons who possess the minimum knowledge and experience necessary to perform tasks on the job safely and competently. The goal of licensure certification is to protect the public from unsafe practices.
Several methods have been developed for setting passing points or cutoff scores, on examinations. Each method employs a different approach. Some researchers (such as Goodwin, 1996; Linn, 1978; Stepheson et al., 2000) prefer to separate passing point methods into two categories: item-centered and person-centered. Item-centered methods involve judgments about the test items. Examples of item-centered methods include the Angoff, Nedelsky, Ebel, and Jaeger methods. Person-centered methods involve judgments about the test takers. Examples of person-centered methods include the contrasting groups method and the borderline group method (Stepheson et al., 2000).

Although each of these passing point methods produces valuable information when implemented correctly, they also have weaknesses that should be fully considered prior to selecting a standard setting method. The discussion below further distinguishes and elaborates on the differences between item-centered and person-centered approaches.

Person-Centered Methods

Person-centered methods of setting passing points are based on evaluations or judgments about the test takers. The method assesses how the examinee’s knowledge, skills, and abilities (KSAs) relate to his or her test scores. This method demands that certain requirements be met for these judgments to be valid. The first requirement is that judgments must be made by qualified judges who are able to determine both the KSAs of an examinee as well as the KSAs that examinees must possess to pass the test. In addition, judgments must reflect the KSAs of the examination in question and not personal characteristics of the test takers. Finally, judgment must reflect the true opinions of the judges (Stepheson et al., 2000).
As mentioned earlier, each standard setting method has certain weaknesses that must be addressed before its implementation. A notable weakness with the person-centered method includes errors that result from inconsistencies between the judges involved in the procedure. Inconsistencies between judges can be attributed to either poor selection or poor training of the judges. Judges who do not possess the KSAs necessary to make appropriate decisions about test takers should be excluded from the procedure. Even when judges have the necessary KSAs, they should be extensively trained so that they fully understand the requirements and implications of each rating. Other weaknesses of the person-centered method are biases that may result from a poorly defined content domain, inappropriate item sampling procedures, and inadequate sample size (Linn, 1978). Two examples of person-centered methods are discussed briefly below: the contrasting groups method and the borderline group method. However, person-centered methods of setting passing points are not the primary focus of this study.

The Contrasting Groups Method

The contrasting groups method is rooted in the theory that judges can divide examinees into two categories of contrasting groups based on the knowledge and skills of the examinees: the performance of qualified candidates and the performance of unqualified candidates. SMEs define satisfactory and unsatisfactory performance on the test after first defining what constitutes unqualified and qualified performance. A pass point is set in the area where qualified and unqualified performance overlap. This method promotes validity by ensuring that performance on the test is directly related to performance on the job (Koffler, 1980; Zieky & Perie, 2006).
The Borderline Group Method

The borderline group method is based on the theory that standards should be set at a point on a measurement scale that would be expected from a barely proficient examinee who has borderline KSAs. SMEs are asked to make judgments about marginally qualified individuals who are in between the highly qualified examinees and the highly unqualified examinees. Using this method, SMEs are first asked to define marginal performance, which is characterized by performance that falls somewhere between satisfactory and unsatisfactory performance. Once a clear definition of marginal performance is agreed to by all the SMEs involved, a group of individuals who perform at the marginal level are identified and selected for participation in the next phase of the process. The performance of the marginal individuals on the examination is recorded and analyzed. A cutoff point is set at the median test score of this borderline group (Koffler, 1980; Zieky & Perie, 2006).

Item-Centered Methods

Item-centered or examination-centered pass point methods require subject matter experts to identify test takers who are proficient enough to answer a test item correctly. SMEs then review items from the test and determine how likely a barely proficient individual would respond to each item correctly. In order to ensure accurate judgments, SMEs review the tasks required to complete a job successfully and the knowledge required to complete those tasks. After reviewing the task and knowledge requirements, and after the definition of the barely proficient candidate is accepted by the judges involved, mathematical estimates are made about the individual test items by each SME, usually on a scale of 0 to 1. The results of all SME judgments are usually compiled using
The Nedelsky (1954) method of determining cutoff scores was founded on the theory that minimally qualified candidates should be able to eliminate obviously wrong answers to a question and then correctly guess from the remaining choices. In the Nedelsky method, a group of judges, usually SMEs, are asked to review test questions independently and then cross out choices that a minimally qualified candidate will be able to identify as wrong answers. The pass point is then estimated to be a function of a chance-guessing proportion of the remaining choices (Skakun & Kling, 1980). This method is mostly utilized with examinations that consist of multiple-choice items because of its reliance on elimination of choices.

As compared to other methods, the Nedelsky method is less concerned with the examinee’s knowledge of the correct answer to the question; rather, it is concerned that the examinee has the necessary knowledge, skills, and abilities (KSAs) to guess correctly. The goal of the Nedelsky method is that examinees will increase their chances of answering the question correctly by eliminating the wrong answers. However, Melican, Mills, and Plake, (1989) cautioned that candidates do not always pick randomly from the remaining choices once the obviously wrong answers have been eliminated. In addition,
Halpin and Halpin (1987) claimed that the Nedelsky method has lower fail rates as compared to other examination-centered methods, and that there are greater variations between rater test scores.

Similarly, Koffler (1980) compared the cutoff scores derived from the Contrasting Groups method of determining cutoff score to that of the Nedelsky cutoff score in an eleventh grade mathematics test that was used by teachers to determine the mastery or non-mastery status of the students. He found that on average, the Nedelsky cutoff scores were lower than those produced by the Contrasting Groups method. He maintained that this is good because the Nedelsky method discriminated between the students who should be classified as masters and those who should not be classified as such (non-masters). Koffler (1980) argued that with the Contrasting Groups method, all the students would have been classified as masters. Furthermore, Glass (1978) contended that in the Nedelsky method, there are low correlations between item minimum pass indices and traditional item difficulty measures.

The Ebel Method

The Ebel (1972) method of determining cutoff scores can be used for any type of testing such as true-false, multiple choice, short answer, or fill-in-the-blank. Ebel (1972) proposed dividing the test items by their level of difficulty (i.e., easy, medium, and hard); as well as their knowledge, skills, and abilities (KSA); and importance (i.e., questionable, acceptable, important, and critical) prior to making judgments about the responses of a minimally competent candidate. Ebel (1972) also suggested that test items should be judged in groups instead of an individual basis.
The Ebel method often requires two reviews before the implementation of test items. According to Cizek and Bunch (2007), during the first round of review in the traditional Ebel method, the focus is on the difficulty and relevance of each test item. When SMEs are assessing the difficulty of the test items, they usually classify the test questions into three categories: easy, medium and hard; and when assessing relevancy, SMEs classify questions into categories such as essential, important, acceptable and questionable (Cizek & Bunch, 2007).

During the second review, SMEs judge each category or classification as a set and a proportion or percentage is assigned to each set. Usually, this step is done as a group. The intention is to achieve a group consensus for each set. The assigned percentage is determined by the proportion of items in a particular set or category that a minimally competent test taker would be expected to answer correctly. Next, a cutoff score is obtained by multiplying each category by its corresponding proportion. The products from the categories are then summed up and divided by the total number of items in that set (Cizek & Bunch, 2007).

The Ebel method is not without criticisms. Ebel himself warned that finding a point on the score scale such as in the midway between the ideal mean score and the expected chance score might be inappropriate if the test items did not discriminate well enough (Ebel, 1972). Similarly, Meskauskas (1976) cautioned that because test items are rated in sets, there might be a higher incidence of judgment errors than those that involve single items.
The Bookmark Method

The Bookmark standard setting method that was developed by Mitzel and Green (1996) is another widely used examination-centered approach. This method also employs subject matter experts to examine items on the test and then give estimation on how a barely proficient student would perform on the items. This method uses test items that have been scored using item response theory. Item response theory is based on the relationship between the difficulty of a test question and the likelihood that the test taker will answer the question correctly (Zieky & Perie, 2004).

In the Bookmark method, items in a test (or a subset of tests) are ordered by their level of difficulty. Judges or subject matter experts are instructed to place a "bookmark" at the point between the easiest items that a barely proficient examinee would answer correctly and the most difficult items that the barely proficient examinee would not be likely to answer correctly. The location of the bookmark point is the cutoff score for the examination (Zieky & Perie, 2004). Since the formation of the bookmark standard setting method, over 28 states have implemented it in setting cutoff scores on their large-scale assessments (Egan, 2002).

The bookmark method uses an Ordered Item Booklet (OIB) and an associated item map. The OIB is constructed using items from the test. The items are ordered in terms of difficulty where the easiest item appears first and the hardest item appears last. The ordering is determined by student performance on the test. The item map provides detailed information about each item in the OIB. Among the types of information included is a ranking of the difficulty of each question, the correct answers to each
question, and the previous location of the test item before reordering took place. It also gives standard setting experts a place to record their observations about each test item (Egan, 2002).

The Bookmark method typically includes training, three rounds of activities and discussions. In the first round of discussion, judges, usually SMEs in small groups of about six to eight examine each item in the OIB and discuss what each item measures and what might be the item’s difficulty level. After discussion, each SME determines a cutoff score by placing a bookmark in the OIB according to his or her own judgment of what test takers should know and be able to do (Egan, 2002).

During the second round of discussion, the judges are encouraged to engage in group discussions about the rationale behind their original scores. During the third round, all of the judges who participated in setting the cutoff scores assemble for a general discussion about their bookmark placements. After each round of discussion, judges may adjust or maintain their bookmark placements. During the third round of discussions, the percentages of test takers who have answered the items correctly are then shown to the judges in the form of proportion data. After the final round of discussion, the test facilitator calculates a cutoff score based on the group’s recommendations by finding the median score from all bookmark placements (Egan, 2002).

One of the advantages of the bookmark method is that it requires less data entry than other judgmental standard setting methods. Also, the Bookmark method takes less time for subject matter experts to finish their tasks. Another advantage of this method is that it is not limited to a specific type of question (Zieky & Perie, 2004).
Buckendahl, Smith, Impara, and Plake (2002) contend that the widespread use of the bookmark method is due to its simplicity and ease of use. Furthermore, the bookmark method does produce more encouraging results in comparison to other standard setting methods (Buckendahl, Smith, Impara, & Plake, 2002).

However, the bookmark is not without criticism. Some researchers question its use of an arbitrary value in establishing the location of the bookmark cutoff score along the scale that is used to rank-order test questions for the subject matter experts’ booklets. Typically, a probability value of success is .67, but many researchers argue that there is no compelling rationale for choosing this particular value over other plausible values such as .50 or .75 (Buckendahl et al., 2002).

The Angoff Method

One widely used standard-setting approach in test development is the Angoff method. It is used in a variety of settings such as education, licensure, and employee selection. The Angoff method is rooted in the theory that a panel of judges, or subject matter experts, can establish a minimum level of proficiency that examinees must meet through setting a passing cutoff score. As described by Angoff (1971), this minimum level of proficiency represents the level of knowledge, skills, and abilities that all test takers must possess in order to succeed in the examination.

The Angoff method relies on subject-matter experts (SMEs) who examine the content of each test question, or item, and then estimate the probability of whether a minimum proficiency candidate will answer the item correctly (Hurtz & Hertz, 1999). In instances where there are discrepancies in the SMEs’ initial ratings of a test item, they are
often given the opportunity to modify their ratings. A mean score is obtained by summing rating estimates of each SME for each test item, and dividing this sum by the number of SMEs. In turn, the mean becomes the passing, or cutoff, score (Ferdous & Plake, 2007).

As mentioned earlier, the Angoff method is one of the most widely used criterion-referenced measures. This is probably due to its systematic approach in deciding on the probability that the minimally acceptable candidate would be able to answer each test item correctly (Angoff, 1971). However, as Hurtz and Auerbach (2003) noted, Angoff did not give any details on how the process of determining the cutoff score should be accomplished. Furthermore, he did not mention how to select or train participants; whether or not to allow participants to discuss choices or revise their judgments; or whether or not participants should be given the answers to the questions that they were judging. Because of the lack of specificity in the original Angoff method, several revisions and modifications have been made by test developers and practitioners. These modified versions are generally referred to as “Modified” Angoff methods.

One of these modifications to the Angoff method is to allow the judges to discuss their estimates with each other after they have provided their initial estimates to the test facilitator. After engaging in group discussion, the SMEs are given a second chance to revise their ratings (Hurtz & Auerbach, 2003). The advantage of this modification is that by allowing SMEs to engage in group discussion, they share information that may help them to make judgments that are more accurate. The disadvantage of this method is that the consensus arrived at by the judges might be influenced by their personal
characteristics or the dynamics of the group that might lead to more inaccurate judgments (Hurtz & Auerbach, 2003). Another modification to the Angoff method is to present normative data containing actual proportion or \( p \)-values of examinees who have answered the items correctly in the past to the judges. The goal of providing normative data to judges is to improve inter-judge reliability (Hurtz & Auerbach, 2003).

The Angoff method does have some weaknesses that are worth mentioning. Because this method involves SME judgments at each step of the process, there is a high potential for human error. Fehrmann, Woehr, and Arthur (1991) warned that the Angoff method displays low reliability between raters and that very little is known about the accuracy of cutoff scores obtained by this method. For this reason, Maurer, Alexander, Callahan, Bailey, and Dambrot (1991) suggested that thorough training should be given to SMEs before the beginning of determining a cutoff score. In the end, using the Angoff Method ensures that the passing grade of a test is determined empirically, which is necessary for a test to be legally defensible and meets the *Standards for Educational and Psychological Testing* (Brennan & Lockwood, 1980).
Chapter 4
APPLICATIONS OF SOCIAL PSYCHOLOGY TO THE STANDARD SETTING PROCESS

As mentioned earlier, all applications in setting cutoff scores in criterion-referenced measures require some sort of human judgment. Chinn and Hertz (2002) asserted that setting the passing score in a licensing examination depends on factors that are specific to the examination as well as to the related profession. They explained that no matter what method is utilized, the process of establishing a passing score often requires exercising judgment of some sort (Chinn & Hertz, 2002). The Angoff method utilizes expert judgment in test development including the setting of a passing score that is based on estimation. The passing score is produced by a team of experienced members in the relevant field of expertise. However, these members may not necessarily be experienced in estimating techniques. Another problem in criterion-referenced measures is that judgments of SMEs vary from person to person. Variability among judges may be greater when it is caused by cognitive biases and social influences. Although the Angoff method is valuable in setting cutoff scores, expert judgment is subject to human biases. Judges’ biases and social influences may lead to errors that could potentially compromise the accuracy of the desired passing score, therefore putting the cutoff score into question.

Cognitive Biases that can Interfere with SMEs Judgment

Although it is very easy for people to make quick and often inaccurate judgments about other people that they barely know, making accurate judgments on something very important such as an examination can be extremely difficult. Shafer and Tversky (1985)
explained that although probability judgment is an element in the process of examination construction, other thought-processes in judges can be induced or evoked that may influence the task of probability judgment. It is possible for cognitive biases to influence judgment in an Angoff passing score setting because SMEs are making judgments about examinees’ abilities to answer test questions correctly. As such, the desired cutoff score accomplished is vulnerable to cognitive biases associated with the standard setting process (Pasisz, 2006). There are several cognitive biases that may influence SMEs judgments.

*The Heuristic Bias*

One of the ways that SME decisions can be influenced is through heuristic cognitive bias. This happens when the probability of an event is assessed based on the availability of data in the memory. Heuristic cognitive bias is a form of distortion in the human mind that can lead to misperception and misjudgment (Jones & Roelofsma, 2000). Heuristics are simplified methods that are put into operation without complete information or accurate calculation. Cognitive heuristic bias stems from humans’ limited information processing capacity that leads to reliance on mental simplifying strategies (Jones & Roelofsma, 2000). This type of bias makes decision makers such as SMEs overlook other important information that may inhibit their objectivity.

Human beings often rely on heuristic methods to assist them in making daily decisions, and SMEs are not exempt. Tversky and Kahneman (1974) noted that the dependence on heuristics and the frequency of biases are not limited to the common person. All heuristic methods (or the peripheral route processing) are vulnerable to
cognitive biases, and have been found by some researchers to bias systematic processing (or central route) methods.

For example, Chaiken and Maheswaran (1994) examined these two routes of processing information in their study using persuasion. The heuristic model contends that several persuasion cues are processed by means of simple schemas or decisions rules. The systematic processing model asserts that persuasion cues are actively attended to and cognitively processed before making decisions. However, Chaiken and Maheswaran (1994) argued that because heuristic cues tend to influence peoples’ expectations about the validity of persuasive messages, they also function to bias peoples’ evaluations of the arguments presented in persuasive messages. Thus, heuristic processing might lead participants to believe that expert messages (rather than non-expert) are valid.

The authors randomly assigned participants ($N = 367$) into small group sessions in which they were asked to read and give their opinions about a new telephone answering machine. The content of the messages contained either weak ambiguous or strong unambiguous information from high credible and low credible sources. The authors found that participants who believed that the product information was prepared by Consumer Reports regarded the message source as having greater credibility ($M = 2.17$) but participants who believed that the product information was prepared by a sales clerk regarded the message source as less credible ($M = .74$). The results indicated that source credibility exercised an indirect persuasive impact by positively biasing systematic processing when credibility was high and negatively biasing systematic processing when credibility was low (Chaiken & Maheswaran, 1994).
In another study, Axson, Chaiken, and Yates (1987) found that participants’ level of motivation was affected by audience response cues. They predicted that participants who had low motivation or a low ability to process information, would rely more on extrinsic cues such as an overhead audience than participants who were high in motivation. Participants (N = 134) were asked to rate a cookie-tasting task after listening to a persuasive debate recorded by a public radio. The results showed that highly motivated participants reported putting more effort into listening to the debate (M = 10.49) and were less influenced by the audience response (M = -2.37) than were low motivated participants, who put in less effort listening to the debate (M = 8.56) and were more influenced by audience response (M = 4.58).

The authors further predicted that participants who were low in “need for cognition” and also low in motivation, versus those who were high in need for cognition and also high in motivation, would be more affected by the audience cues. The authors divided the participants into two groups of high versus low need for cognition using a median split. As predicted, low need for cognition participants in the low motivating conditions agreed less when audience response was less enthusiastic (Axson, Chaiken, & Yates, 1987). The results of these studies were consistent with that of other studies in the heuristic bias arena in concluding that heuristic bias does influence people’s decisions. This applies to SMEs as well.

The Insufficient Anchor and Adjustment Bias

Insufficient anchor and adjustment bias typically occurs during estimation. Estimation begins by searching and retrieving from memory other examples that are
similar in ways that may guarantee plausible estimates for the quantity sought. Basically, when people are trying to estimate an unknown quantity, they begin with information that they know, such as using the first reasonable number that makes sense, as a starting point or an anchor for an initial estimate. When re-estimating that effort, insufficient adjustments are made from that anchor so that an acceptable (but biased) value is achieved (Epley & Gilovich, 2006).

Insufficient anchor and adjustment bias is often observed in estimation. When people make their initial decision based on insufficient information, and even after re-evaluation of their initial source of information due to inaccuracy, their re-estimation often does not deviate much from their previous estimates or anchor. Insufficient anchor adjustment bias usually occurs when a decision maker observes a situation that is comparable to a past event. This comparison makes him or her think that a similar outcome will be obtained (Epley & Gilovich, 2006).

Insufficient anchor and adjustment bias can lead decision makers to make a subjective and inaccurate decision. For example, during a new student orientation, freshmen students may be asked to give an estimate of the population of Sacramento State University. Many students might think of the size of other Universities, especially those near their hometowns or cities as a reference, and form estimates from this pre-conceived impression. After stating their estimates out loud, students may be asked to get into discussion groups of five and then come up with their final estimates. According to this paradigm, the freshmen would adjust their estimates after discussing with each other; however, their estimates would probably not differ much from their initial value.
This example is consistent with other studies in this area. For example, Epley and Gilovich (2001) conducted a study in which they predicted that participants \(N = 50\) would describe a process of anchoring and adjustment only when anchors were self-generated. They found that participants adjusted less from values that they had generated as a starting point \(M = 73.9\%\) than when considering experimenter-provided anchors \(M = 13.0\%\). Even when these starting values were proved incorrect, participants’ final values were closer to their self-generated anchors. Similarly, in their meta-analysis, Hurtz and Auerbach (2003) found that when judges were allowed to define the procedure for minimum competence, and when the same group of judges were given the chance to discuss and revised their estimates, a higher degree of consensus was reached as well as a higher cutoff score. This finding can be related to the anchor and adjustment bias in that SMEs who defined the procedure for minimum competency in an Angoff passing score workshop would be more likely to stick closer to the anchor they set, whereas if the definition was given to the SMEs, they would be more likely to deviate from that anchor.

Insufficient anchor and adjustment bias can occur in an Angoff passing score workshop because SMEs are deciding and providing estimates or the probabilities that examinees are going to get an item correct. According to this paradigm, SMEs final ratings would not deviate much from their initial ratings. For example, in an Angoff passing score workshop, a group of six SMEs may rate an item in terms of its level of difficulty as 70, 70, 85, 75, 75, and 80, stating that they (SMEs) do this task everyday at work, and as such, test takers should know it. However, the test facilitator points out that in actuality many test takers did not get the question correct. The proportion of previous
test takers ($p$-value) who got the item correct was just .35. After a group discussion, SMEs decide to change their ratings .10 - .15 points downwards towards this value. In this case, based on the anchor and adjustment bias, a downward revision of .55 -.70 would be expected in their final ratings. Notice that even though SMEs changed their ratings, their final ratings are still anchored from their initial ratings. The observed level of adjustment made is probably insufficient given the wide discrepancy between SMEs’ ratings and the $p$-value. In order words, the final cutoff point that SMEs set for a particular examination may be influenced by the insufficient anchor and adjustment bias.

*The Over Confidence Bias*

Over-confidence bias typically occurs when experts or decision makers are overly optimistic and very confident in predicting the future; they overestimate their abilities and their knowledge when facing complex tasks and underestimate the effort involved in completing the difficult task. Research seems to suggest that generally people tend to be over confident and even more, investors tend to be overly confident in their abilities to predict the future (Daniel, Hirshleifer, & Subrahmanyam, 1998; Odean, 1998).

The more difficult and uncertain a task is the more prevalent is the risk of the overconfidence bias. Overconfidence bias can be commonly observed in investors who attribute favorable market returns to their own personal skills and decision making abilities. In particular, this can be observed in investors who predict the future of companies that are growing fast. For example, Scott, Stumpp, and Xu (2003) tested the theory that over confident investors cause a greater lagged response to news for rapidly growing stocks than for slowly growing stocks. They utilized data from France, United
Kingdom, Germany, United States, and Japan between the years 1987 to 2000. Their results indicated that slowly growing cheaper stocks tended to outperform rapidly growing expensive stocks for all five countries (Scott et al., 2003).

While investors are often overly confident when investments are doing well, they tend to avoid taking responsibility for market declines. Instead they often blame declines in the markets on result of external markets or environmental factors. In simplest terms, overconfidence bias suggests that many investors over assess their own skills. While confidence in investing can be considered an advantage, overconfidence can cause investors to miss the opportunity to learn from past failures.

Similarly, Tversky and Kahneman (1974) presented participants with many paragraphs in which the performance of student teachers was described. They asked one group to evaluate in percentile form the quality of the performance described in the paragraph; and another group to predict also in percentile form what the standing of the teachers would be in five years. They found identical extreme results in both the evaluation and the prediction of the teachers.

Professional examination development is not an easy task and certainly making crucial decisions like setting cutoff scores for an exam that can affect the future of many people can be extremely stressful for those involved. This can happen in an Angoff passing score setting when SMEs who overvalue their profession or SMEs who may want to maintain the reputation of their institutions, may decide that licensure examination based on minimal competency is not the best way to earn licensure certification. Due to their own over confidence, SMEs may perceive aspects of their profession to be easy, and
therefore might assume that everyone should know what they (SMEs) know. As a result, they (SMEs) may rate the test items as too easy thereby calling into question the validity of the examination.

*The Insensitivity to Base Rates Bias*

Insensitivity to base rates bias typically occurs when decision makers base their decision on qualitative information and ignore quantitative information. Kahneman and Tversky (1974) demonstrated this bias in a study in which they presented participants with descriptions of people who came from a fictitious group of 30 engineers and 70 lawyers. The participants were asked to rate the probability that the person described was an engineer. Although the probability of a person being an engineer was just 30%, participants assessed higher probability by matching the description of a person to their stereotype of what an engineer should look like. This means that participants’ assessments were affected by the person’s description and not the information presented.

A similar phenomenon can occur in an Angoff setting where SMEs are asked to read and evaluate scenarios in examinations and then provide ratings to the question(s) related to what is in the scenarios. If the scenarios are related to topics in their area of expertise, they may be biased in their ratings depending on their own familiarity with the content of the scenarios. For example, if a scenario consists of something that they do every day at work, they may rate the probability of examinees knowing the test items related to that scenario as high, and therefore rate the items as easy. Or, they may rate the test items of the scenario as too difficult if it consists of something they encounter only infrequently at work.
The Halo Effect Bias

The halo effect bias is a cognitive bias whereby the perception of a particular trait is influenced by the perception of the former traits in a sequence of interpretations. Thorndike (1920) was the first to demonstrate the Halo effect in empirical research. In his study, Thorndike asked commanding officers to rate their soldiers. He found a high cross-correlational relationship between all positive and all negative traits.

Similarly, Nisbett and Wilson (1977) explored the question of people’s awareness of the halo effect in their study. They argued that people have very little awareness of the type or even the existence of the cognitive processes that intercede with judgments, inferences, and the creation of complex social behavior. They showed two different videotapes of an instructor who spoke English with a European accent to 118 undergraduate students enrolled in introductory psychology. In one of the videotapes, the instructor was warm and friendly. In the other video, the same person was cold and distant. They asked the students to rate the instructor in terms of his accent, appearance, and mannerisms. Their results showed that the majority of the students’ who saw the warm and friendly video, rated the instructor’s mannerisms as appealing, while a majority of the students who saw the cold and distant instructor rated his mannerisms as irritating. Approximately half of the warm condition students rated the instructor’s accent as appealing and half as irritating, whereas the majority of students in the cold condition overwhelmingly rated his accent as irritating.

These results were consistent with previous results regarding the halo effect. It indicated that global evaluations of an individual can generate altered evaluations of the
individual’s attributes, even when there is enough information to permit for independent assessment of them (Nisbett & Wilson, 1977).

In another study, Nisbett and Wilson (1977) demonstrated that generally, people are oftentimes unaware of the existence of a stimulus that influenced their responses. Not only were they unaware of the existence of the stimulus, but they are unaware that the stimulus had affected their responses. Nisbett and Wilson argued that when people attempt to report on their cognitive processes, they do not do so based on any true introspection. Their reports were instead based on judgments about the extent to which a specific stimulus is a plausible cause of a given response. Their results indicated that even when people were told that their judgments have been affected by the Halo effect, they still had no clue as to when the halo effect influenced them (Nisbett & Wilson, 1977).

SMEs and test panelists in an Angoff passing score workshop make important decisions about the cutoff point of test items, but their decision-making processes can be greatly influenced by the halo effect without their knowing. This is because some SMEs may come to passing score workshops with an overall general impression of the test’s level of difficulty. For example, a group of eight SMEs in an Angoff passing score workshop may rate a test item as 95, 90, 90, 95, 85, 90, 95, and 95 indicating that the examination items are all very easy. Or SMEs experiencing a reverse halo bias may rate the test items as 25, 25, 35, 30, 35, 25, 30, and 35 indicating that the test items are very difficult. During group discussions, they insist that their ratings reflect the proportion of candidates who would get the items correct and therefore refuse to change their final
ratings. This overarching common factor may influence all their ratings so that regardless of the actual profile of the test items, a similar rating is given to all test items. Although SMEs rated the test items as easy, in actuality they are not, and so this produces the halo effect.

In a way, the halo effect is somewhat similar to the anchor and adjustment bias because SMEs do not also deviate much from their initial ratings even after group discussions. This type of bias is also not good for setting standards for licensing examinations because a cutoff point for an exam that is set too easy may lead to a situation where unqualified candidates become licensed to practice. In contrast, a cutoff point that is set too difficult might disqualify the minimally competent candidate from being licensed. The halo effect bias is also not good for setting standard because it will undermine the established cutoff scores. This raises the question as to what an examination facilitator can do to minimize biases of SMEs in an Angoff passing score workshop that may be caused by the halo effect and other cognitive biases.

The Central Tendency Bias

In descriptive statistics, Reaves (1992) described the central tendency as the middle value or a representative value of any probability distribution or a set of scores. It is the tendency of samples of a given measurement to cluster around some central value. However, central tendency bias refers to the tendency of raters to evaluate their ratings in the middle when they apply a rating scale. For example, given a scale with points on it that run from one (poor) to seven (excellent), with four being the middle, many raters will refuse to use the points at either of the ends.
The difference between the central tendency bias and the anchor adjustment bias is that SMEs who are experiencing the central tendency bias tend to rate all the test items in the middle of the scale before and after group discussions. On the other hand, in the anchor and adjustment bias, SMEs would utilize the entire scale for their ratings but they would not deviate much from their initial ratings even after group discussions. This can be problematic in an Angoff passing score setting because the cutoff point for that exam may be set in such a way that unqualified candidates who should not be allowed to practice will become licensed. Furthermore, professional exams need to be fair, valid, and legally defensible. The prevalence of central tendency bias would cause the exam to be unfair and therefore invalid to meet examination standards.

Social Influences that can Interfere with SMEs Judgment

In an Angoff passing score setting, SMEs are given the chance to read and then rate each item on a specific exam related to their field of expertise. SMEs are usually given a second chance to discuss their ratings and then revise their initial ratings as needed. Hambleton and Plake (1995) explained that the modified Angoff method can involve up to three rounds of ratings due to variability in judges’ estimates. The ratings for each item are then summed and averaged to have the cutoff score for that particular examination. The goal of having additional rounds is to reduce variability in judges’ ratings (Hambleton & Plake, 1995).

However, some researchers argued that although group discussion is important because it helps in reducing interjudge variability in ratings, some SMEs could be subjected to social influences during these group discussions that may create pressure to
make rating decisions that they had not intended (Fitzpatrick, 1989; Hurtz, 1998). Group discussions in an Angoff passing score setting have been shown to be beneficial and valuable because they allows for an exchange of information between experts (Hurtz & Auerbach, 2003).

Several studies have found that group discussions do in fact reduce variability of item estimates across panelists. For example, in their meta-analysis, Hurtz and Auerbach (2003) found that not only does group discussion lead to reduction in variability among judges but group discussion also led to a higher cutoff score. In other words, although group discussion is valuable in an Angoff passing score setting, it is susceptible to social influence factors that might in turn influence how SMEs rate test items. Jones and Roelofsma (2000) described social influence as “the process by which individual judgments, behavior, and attitudes change as a result of the real or implied presence of other people” (p.1133). Social influence bias stems from group discussions, and as such it is regarded as discussion-based (Jones & Roelofsma, 2000).

Kerr (1981) utilized a path independence model to investigate whether the likelihood of a group shifting its position is independent of all previous states of the group, (except the state that it is presently in) and if movement toward agreement on a popular view was more evident in groups of women or men. Participants consisted of 282 undergraduate students who participated in the study as a partial fulfillment of a course requirement. The results of his analysis indicated that during the group decision process, group members were more likely to join a majority than to defect from one, ($p < .01$) and that movement in one direction signaled a greater likelihood of further
movement in that direction. In addition, he found that women were less likely to voice their opinion but were more likely to voice their support for others than were men.

This study illustrates that social influences affect people’s behavior in a group setting. Chinn and Hertz (2002) warned that whenever group discussion is allowed in a passing score setting, it becomes more difficult for judges to make their ratings truly independent of their own judgments. This is particularly true within modified Angoff passing score settings because SMEs typically need to show an acceptable level of agreement before the cutoff point is established. The goal of SMEs in an Angoff passing score workshop is to achieve consensus on test items so that a passing score for that specific examination is obtained.

Conformity Pressure

In many group situations in which an important group decision needs to be made, there are often other processes and stimuli such as anxiety and biases taking place that might interfere with the decision made. This is because groups have a pervasive and substantial impact on the behavior and attitudes of individual group members. Groups can powerfully affect how people think and act in a working environment. The attributes of the people in the group and the nature of the task can influence behavior of other group members forcing them to reach decisions that are contrary to what they actually think. As explained by Epley and Gilovich (1999), conformity pressure often generates a hard feeling of internal conflict that makes an individual feel obligated to cave in even though he or she knows that the attitudes, opinions, or behaviors of others are wrong. Aronson
(2004) defines conformity pressure as “a change in a person’s behavior or opinion as a result of real or imagined pressure from a person or a group of people” (p. 16).

Beloff (1958) distinguished between two types of conformity: acquiescence and conventionality. He described acquiescence as the type of conformity that results from pressure from other people with different opinions. In many experiments illustrating this type of conformity, it is usually arranged so that a personal social influence would invoke inconsistency (or raises doubts) with previously elicited responses that would tend to distort participants’ judgments. He described conventionality as the agreement with the ideology, attitudes, and belief of a person’s culture or subculture (Beloff, 1958). The pressure to conform usually can be seen in group members who hold different views within a group. Group members may pressure those with different views into revisiting or revising their positions so that a consensus can be reached within the entire group. This was beautifully displayed in Asch’s (1956) experiment in which he asked participants to say which line from a set of three straight lines ABC was closer to another line X. Although line B was clearly the correct answer, participants were pressured to conform with other members even though they knew the correct answer.

In another study, Epley and Gilovich (1999) conducted an experiment in which they investigated whether priming the construct of conformity or non-conformity can lead people to be more inclined to follow or resist the influence of others. They presented a task made up of strings of five scrambled words to thirty-four undergraduate students and asked them to use four of the words in each string to create a grammatically correct sentence. The experimenters included a priming word that was semantically related to
either conformity or non-conformity to the different condition groups. The experimenters found that participants primed for conformity rated the task more favorably and expressed views that were similar to those of experimental confederates than those primed for non-conformity.

Also, in an experiment, Torelli (2006) asked introductory business students ($N = 89$) to read a story that was aimed at priming either their conception of self as connected to others (interdependence self-concept) or their conception of self as separate from others (independent self-concept). He further asked his participants to perform an unrelated task in evaluating products and to then explain their judgments about the products to other participants in small group discussions. The results showed that students whose interdependent self was primed were more likely to conform to others opinions, especially when they anticipated explaining their judgments to others. Those whose independent self were primed were less likely to conform.

In another study, Moscovici and Personnaz (1991) showed slides of different colors that consisted of lines that eventually made a picture of Vladimir Lenin to thirty female students in four phases. The authors found that when the participants were faced with a majority opinion, they often compared their responses and worried whether their responses agreed with that of the group. Often, this comparison led them to comply with the majority judgment. This result demonstrates that the pressure to conform and follow the ideas of others has an influence on individual behavior.

In an Angoff passing score workshop, the pressure to conform can be enormous because SMEs oftentimes have to reach a consensus on their final ratings for each test
item before moving on to the next item. Typically, group discussions are often carried out whenever SMEs’ ratings are out of range. During these group discussions, there is often pressure for SMEs with extreme ratings to reconsider their ratings so that a consensus can be reached. The pressure to conform is particularly more intense when the $p$-value of previous examinees is shown to the SMEs. For example, a group of six SMEs in an Angoff passing score workshop may give their initial ratings as: 45, 70, 75, 80, 75, and 85. Notice that the range of this group is 40 (i.e., the lowest rating and the highest rating, $85 - 45 = 40$). During the group discussion, the two raters with extreme ratings (i.e. 45 and 85) would be asked to give the rationale for their ratings. It is during group discussions that other SMEs might pressure those with extreme ratings to increase (or decrease) their rating so that a consensus can be achieved. After group discussions, SMEs are then asked to give their final ratings. The revised ratings have a tendency to move towards the group’s central tendency if conformity pressure is the social influence factor being observed (Hurtz, 1998).

Group Polarization

In addition to both minority influence and conformity pressure, another social influence factor that can occur in an Angoff passing score setting is group polarization. This type of social influence is defined as the tendency for a group’s decision as well as the mean of the individual opinion positions held by the group member, to become more extreme after social interaction than before interaction (Fitzpatrick, 1989; Hurtz, 1998; Myer & Lamm, 1976).
Fitzpatrick (1989) explains that polarization can be observed if the group’s initial opinion shifts to a more extreme position after group discussion. Furthermore, Hurtz and Auerbach (2003) noted that group polarization occurs when the initial position of a group member who was moderately leaning towards the high or low end of the rating scale, becomes even more extreme towards a specific direction after group discussions of their position. Unlike the conformity pressure where the revised ratings are tended towards the group’s central tendency, with the group polarization factor, the group will become more extreme towards whatever direction the group was originally leaning. This could either be towards the low or high end of the measurement scale. This phenomenon is basically saying that when a group converges on a position after a discussion, the convergence will be more extreme than members’ original positions.

Myers and Lamm (1976) described two types of group polarization: the risky shift and the cautious shift. The risky shift occurred when the overall group becomes more risk seeking than the average initial risk seeking behaviors of the individual members. The cautious shift occurred when the overall group became more risk averse than the average risk averse tendencies of the individual members. In both types of shift, the average response of individuals in the group becomes more extreme after group discussion. Several studies have demonstrated these risk phenomena.

For example, Myers, Schreiber, and Viel (1974) asked felons in a correctional institution ($N = 66$) and male students ($N = 67$) to respond to ethical-legal dilemma issues. They predicted that there would be a shift to increase in risky behavior when the participants initially perceived themselves to be riskier than the average subject. The
authors used scenarios in which an older teenager was contemplating whether to buy a drink or not. A Likert-type response scale was utilized. The results showed that after group discussions, there was a significant increase for both the students and felons who initially liked the behavior of the teenager. In addition, participants in a study conducted by Horne and Long (1972) engaged in group discussion about situations involving conflicts between particularistic and universalistic moral obligations. They found that participants who initially valued universalism, shifted towards increased universalism after group discussion. The overall behaviors of the participants in these studies after engaging in group discussions, illustrate the group polarization effect.

To further demonstrate the group polarization effect, Myer and Kaplan (1976) developed a simulated jury setting with social psychology students (N = 60) and asked the students to first respond to eight hypothetical traffic cases which differed in their implication of guilt (high or low). Then, the authors asked the students to discuss the cases in small groups and respond a second time to all eight hypothetical cases. As the authors predicted, the simulated jury deliberations polarized the mean judgment of the cases after group discussions. After the students held discussions for low guilt cases, they were, on the average, more extreme in their judgments of innocence and more lenient in recommended punishment. Similarly, after group discussion for the high guilt cases, their decisions shifted towards harsher judgments of guilt and punishment. However, these effects were not observed in cases in which no group discussions were held (Myer & Kaplan, 1976). Several mock jury findings (e.g., Bray & Noble, 1978; Sunstein, 2007) have supported the group polarization paradigm.
It is very common for SMEs in an Angoff workshop to engage in group discussion. As Fitzpatrick (1989) noted, convergence of opinion has been noticed after group discussion in several situations where pre- and post-discussion variability were compared. The findings above seem to suggest that SMEs who held extreme opinions and initially rated an item as either hard or easy, may become more extreme after group discussions. Thus, their final ratings would also become more extreme.

The impact of obtaining a group consensus may undermine the acquired cutoff score. Allowing SMEs to have group discussions after their initial ratings and giving them a chance to revise their ratings, could lead to a reduction in variability among judges, and the establishment of higher cutoff scores (Hurtz & Auerbach, 2003). As a result of group polarization, SMEs’ ratings may lead to either a situation where the barely competent candidate is denied license (false rejection) or a situation where the unqualified candidates are licensed (false acceptance).

**Minority Pressure**

In many group discussions, members who hold minority views are often persuaded to join the views of the majority so that a consensus is reached. Several studies in the social influence arena have shown that when a viewpoint is given by a minority source, the majority of the individuals are not motivated to adopt it (Moscovici, 1980). However, there are a few situations in which minority members succeeded in persuading the majority to reach a consensus. Minority influence occurs when an individual group member successfully persuades the majority members of the group to accept his or her viewpoint. According to research, minority influence can be successful when there is
consistency in the minority message or when the minority position is different from that
of the norm (Gordijn, de Vries, & de Dreu, 2002).

This paradigm was demonstrated in a study by Moscovici and Personnaz (1980)
in which they asked participants to view a set of slides and state their colors. The
participants consisted of six participants (the majority) and a confederate (the minority).
Although all slides were blue, the minority consistently insisted that the color of the
slides were green. The results indicated that among those subject exposed to the minority
view, 32% reported seeing green at least once.

In another study carried out by Gordijn et al. (2002), they conducted three
experiments in which they investigated changes in minority position. In experiment 1,
they sought to understand if a change in minority size motivates people to think of the
minority position in terms of their attitudes on the focal issue and their attitudes on the
related issue. The participants consisted of 55 students who were randomly assigned to
the conditions of a 2 (argument quality: strong vs. weak) X 2 (minority size: expanding
vs. shrinking) factorial design. The results did not show any minority influence on the
focal issue. This was consistent with past findings. However, the result on related issues
was significant with a main effect for argument quality, and the predicted Argument
Quality X Change in Minority Size interaction. In the case of an expanding minority,
strong arguments led to a more positive attitude on related issues than weak arguments.
The authors found similar results in experiments two and three investigating the same
variables (Gordijn et al., 2002).
To further illustrate the minority influence factor, Sinaceur, Thomas-Hunt, Neale, O’Neil, and Haag, (2010) examined whether a minority faction that was perceived as expert would enhance the quality of decisions made by its group members. The authors predicted that perceived expert minority as compared to non-expert, would make majority individuals more accurate in their private decisions after group discussion. This improved accuracy would be seen as a result of greater questioning of the validity of the majority’s arguments. The authors further predicted that minority members’ perceived expertise would make them less likely to change their own decisions. The authors tested these predictions in two experiments. Based on a sample made up of graduate students ($N = 235$) who were randomly assigned to 47-five-person groups, they found that majority members held more accurate private judgments after group discussions with perceived expert minorities, even though publicly, the group decision was not more accurate. Also, minority members who believed in their own expertise were less likely to change their private judgments. Similar results were found across studies (Sinaceur et al., 2010). This study supports the notion that minority opinion can sometime influence majority judgments.

Also, in two experimental studies, De Dreu and West (2001) found that in organizational team decision making processes, minority dissent often led to high levels of innovation and divergent thought. Based on a sample ($N = 109$) consisting of 21 teams made up of about five members per group, they predicted that minority dissent in organizational teams would increase creativity if team members participated in high (and not low) levels of group decision-making process. Their results were consistent in a
homogenous set of teams sampled from the national postal service and also in a heterogeneous set of teams from several organizations (De Dreu & West, 2001).

Furthermore, Kerr (2002) found that messages with relevant information were processed systematically when the source was from an active minority and not from a passive minority. This finding was consistent with other scholars (e.g. De Dreu and West, 2001; Gordijn et al., 2002) who have affirmed that when an issue was not of high relevancy and it was being advocated by the majority members, it was generally accepted by the targeted audience even when the argument presented was weak. However, if the issue was important enough, people paid more attention when the advocate was a minority (Kerr, 2002).

As evident in these studies, there are occasions when a minority view may influence that of the majority. It all depends on the consistency of the minority and his/her style of argumentation. In an Angoff passing score setting, this can occur during group discussions to reach a consensus; and especially, when the minority is perceived to be an expert on a particular content area. He/she may be able to convince other subject matter experts that the test items in a given area relate to his/her daily practice.

For example, in a veterinary medical workshop, other SMEs who do not specialize in the treatment of large animals such as cows might adopt the views or position of a Vet doctor who specializes in treating large animals, and then revise their initial ratings towards that of the minority opinion. Moreover, in his experimental study consisting of undergraduate students (N = 44) and a confederate minority deviant, Pasisz (2006) predicted that as a result of exposure to the minority deviants, judges in an Angoff
standard setting workshop would revise their ratings in the direction of the minority. His findings supported this prediction. Minority deviants can influence judges to revise their ratings towards the direction of that of the minority. Furthermore, an SME with a strong opinion about his/her extreme rating may refuse to revise it during group discussions because he/she believes that the test item is very difficult (or very easy). In this case, it is up to other SMEs to revise their ratings toward this individual so that a consensus can be reached. This behavior would totally undermine the expertise of the other SMEs and the validity of the probability estimates given by the SMEs. This in turn undermines the cutoff score that has been established.
Chapter 5
THE CURRENT STUDY

The purpose of this study is to investigate evidence of social influence factors and cognitive bias in the Angoff method of setting passing scores through the analysis of pre- and post-discussion rating data. The appropriate way to set cutoff scores in testing and determining competency is a subject of great debate in the world of testing. While the Angoff method is commonly used, standard setting methods in general are subject to preconceived judgments and more importantly, cutoff scores are the product of subjectivity by the judges involved. In particular, it is not well understood how social factors influence the decisions that judges make in setting the passing scores of examinations. As McGinty (2005) notes, research is still lacking in knowing what goes on in the minds of judges.

This study will investigate how social influence factors and cognitive bias influence the decisions of SMEs in Angoff passing score workshops. Conformity pressure, group polarization, and minority pressure are the social influence factors that will be investigated. Anchor and adjustment bias is the cognitive bias that will be investigated.

Although researchers have conducted several studies emphasizing that high levels of cognitive processes occur during judgmental standard setting tasks or workshops, few researchers (e.g., Fitzpatrick, 1989; Hurtz, 1998; McGinty, 2005) have looked into how social and cognitive influences may influence SMEs ratings. This study investigates how social influences may have guided the judgments of SMEs by comparing their initial and
final ratings in setting cutoff scores for several licensure examinations. Archival data from several Angoff passing score workshops were analyzed for evidence of social influence effects and cognitive bias. More specifically, the study looked at changing patterns between the initial and final ratings of SMEs. This study focused on whether there is evidence that there may be group discussion effects that might have influenced SMEs’ decisions in revising their final ratings, and whether group processes might have influenced their decisions in setting the final cutoff scores.

Observable Patterns of Conformity Pressure, Group Polarization, Minority Pressure, and Insufficient Anchor and Adjustment Influences

In order to distinguish the effects of the above-mentioned paradigms, it is important to identify the outcomes that would result from these four paradigms. The expected observable patterns will be the direction of change in the ratings or the probability estimates given by the SMEs for each test item, and the variability in their estimates. The direction and variability in SMEs ratings would determine which of the paradigm discussed earlier might be evident in Angoff passing score workshops.

Patterns of Conformity Pressure

The conformity pressure paradigm emphasizes feelings of internal conflict that make an individual feel obligated to cave in even though he or she knows that the opinions or behaviors of others are wrong. This paradigm predicts that SMEs with extreme initial ratings would revise their ratings towards the central distribution of the measurement scale. Group discussions in an Angoff passing score workshop are often carried out when SMEs initial mean is out of range (i.e., the difference between the
highest rating and the lowest rating is more than 25). This may create pressures for SMEs with extreme ratings to revise their ratings so that a consensus is reached. The presence of conformity pressure would then be indicated by inward revision of extreme ratings from both ends of the measurement scale (i.e., .25 -.95) towards the group’s central tendency. The expected observable patterns in the data will be the differences in standard deviation. If the standard deviation moves inward, that would be an indication that the standard deviation is shrinking and a reduction in variance would therefore be expected. The means, however, are expected to stay the same.

Patterns of Group Polarization

The group polarization paradigm predicts that SMEs who had rated an item as moderately difficult in their initial ratings, may revise their ratings after group discussions to be more extreme towards the direction that the group tended to agree to (Hurtz, 1998). Support for the group polarization paradigm can be observed if the group’s initial opinion shifts to a more extreme position after group discussions. For example, if the probability estimates provided by a group of SMEs in an Angoff passing score workshop ranges from a high of .75 to .95, it would be expected that SMEs who rated the item closer to .75 on the measurement scale, would revise their probability estimates upwards closer to .95. A more extreme rating due to group polarization would indicate that the item was easy after group discussions. Similarly, if the probability estimates ranges from a low of .45 to .75, SMEs who rated the item closer to .75, would be expected to revise their estimates downwards closer to .45, a more extreme rating, due to group polarization, indicating that the item was challenging after group discussions.
(Fitzpatrick, 1989; Hurtz, 1998; Hurtz & Auerbach, 2003). The expected observable patterns in the data will be changes in means and reductions in standard deviations towards the direction that the group was originally leaning.

Patterns of the Minority Pressure

The minority influence paradigm predicts that an individual group member can successfully persuade the majority members of the group to accept his or her viewpoint. This paradigm contends that minority influence can be successful either when there is consistency in the minority message or when the minority position is different from that of the norm. Support for this paradigm can be observed in the data if there is a shift in the final estimates of the majority members from their initial estimates towards that of the minority. This shift can be observed from both ends of the distribution of the scale. SMEs who initially rated an item as .45 (as difficult) may move up slightly towards the position of the minority; or SMEs who initially rated an item as .85 (easy) may move down slightly towards the position of the minority. The expected observable pattern in the data would be indicated by the differences in means of SMEs final ratings.

Patterns of Insufficient Anchor and Adjustment Bias

The insufficient anchor and adjustment bias paradigm predicts that during estimation, SMEs would retrieve information that they know from memory and utilize that information as an anchor for an initial estimate. However, during re-estimation of their final ratings, even after group discussion, insufficient adjustments would be made from their initial anchor so that an acceptable value is not achieved (Epley & Gilovich, 2006).
Support for this paradigm can be observed in the data by computing each SMEs means and standard deviation across all items rated, and then comparing how much variation there is across SMEs. It is expected that SMEs would anchor their ratings within a narrow range around .70 supporting a common folklore that surrounds the Angoff method in setting cutoff scores (Hurtz & Auerbach, 2003). In addition, it is expected that SMEs’ final ratings would not deviate much from their initial ratings.

The four variables mentioned above are the independent variables for this study and the dependent variables were the direction and amount of change in probability estimates given by the SME before and after group discussion. It is hypothesized that across a large number of SME panels, there will be instances of each of the social influence/cognitive bias processes being evident in the data. Specifically, it is hypothesized that:

1) There would be evidence that SMEs with extreme ratings on either end of the measurement scale would revise their estimates inward toward the central tendency of the group, suggesting the presence of conformity pressure.

2) There would be evidence of group polarization effect in that the mean ratings from pre-to-post discussion would become more extreme while variability among group members would at the same time decrease.

3) There would be evidence of a single SME influencing others during group discussions, such that a shift will occur in their final ratings towards that of the minority opinion holder.
4) There would be a consistent mean across raters and the judgments SMEs provide would not deviate much from that mean, suggesting an insufficient anchor and adjustment bias.
Chapter 6

METHOD

Research Setting

The Department of Consumer Affairs (DCA) is a state government agency in California that is devoted to the protection of its consumers from unfair competitions from unlicensed practitioners (www.dca.ca.gov). DCA currently has 39 regulatory entities including twenty-five boards, ten bureaus, seven DCA programs and committees, and fourteen offices and divisions that regulate over 100 businesses (www.dca.ca.gov). The Department of Consumer Affairs issues licenses for more than 200 professional categories. Some of these categories include physicians, architects, embalmers, contractors, cosmetologists, accountants and automotive repair persons (www.dca.ca.gov). The goal of the DCA’s licensing entities is to “establish minimum qualifications and levels of competency for licensure” (DCA-Office of Publications, Design, & Editing, 2008).

The Office of Professional Examination Services (OPES) is one of the divisions within the Department of Consumer Affairs that provides “examination-related services” to different boards and bureaus within the State of California. Examples of these boards include the Board of Acupuncture, the Board of Court Reporters, the Board of Behavioral Sciences, the Board of Psychology, and the Board of Optometry just to name a few. Examples of the bureaus include the Bureau of Automotive Repair, the Bureau of Cemetery and Funeral, and the Bureau of Security and Investigative Services (DCA-Office of Publications, Design, & Editing, 2008).
Some of the examination-related services provided by the Office of Professional Examination Services to the Department of Consumer Affair’s regulatory entities include performing occupational analyses or job analyses that “define a profession in terms of the actual tasks that new licensees must be able to perform safely and competently at the time of licensure” (OPES-Informational Series, 1, 2003). OPES’ goal is to develop a licensing examination that is solidly based on what licensees actually do on the job and that the content of the examination is “fair, job-related, and legally defensible” (OPES-Informational Series, 1, 2003).

The Office of Professional Examination Services also engages in conducting examination item development. The intention is to protect consumers by making sure that new licensees possess the minimally acceptable knowledge and skills needed to perform tasks on a specific job safely and competently (OPES-Informational Series, 3, 2003). OPES utilizes the test scoring and item analysis methodology to score and evaluate the effectiveness of each test question within an examination (OPES-Informational Series, 2, 2003). Furthermore, OPES provides psychometric expertise and prepares validation reports required for the development of professional examinations in California.

Test Content

This study utilized test items from several licensure examinations. The names of all examinations were kept confidential and instead numerical values were utilized to specify the different examinations in this study. The examinations comprised of multiple-choice test items that cover all content areas of the professions chosen. The test items were developed using relevant tasks and knowledge statements from the examination
blueprints that were derived from the relevant occupational analyses. The test items were selected from test banks. Items selected were based on their quality of past administration (\(p\)-values). This study included data from 38 different Angoff passing scores workshops, comprised of 3,983 test items.

Judges

The judges were subject matter experts from several professions. There were between five to eight SMEs in each passing score workshop, making a total of 260 SMEs that are included in this study. Numerical values were assigned to replace names of judges. All judges’ information was kept anonymous and confidential and was treated according to APA ethical standards.

Standard Setting Procedure

Subject matter experts (SMEs) for the chosen professional examinations were invited to attend a passing score workshop at OPES. Arrival time and days are usually pre-arranged, same as the length of time allocated for the workshops. At OPES, in order to maintain security of the test items, the photo identity cards of SMEs are checked to make sure that they are who they say they are. The test facilitator then directs them to one of the examination construction rooms. The test facilitator welcomes them and he or she introduces her/himself. The test facilitator usually asks SMEs to introduce themselves as well, stating where they are from (location) and how long they have been practicing in their profession. The test facilitator would explain the rules in workshops and the consequences of compromising the examination. Several security forms are explained and then given to SMEs asking for their signatures. Name badges are often given to the
SMEs and they are required to have it on at all times while they are in the testing area. SMEs are usually asked to turn off their cellular telephones and put their private belongings away from the test area. Two 15 minutes breaks and a one hour lunch break are often given to the SMEs. However, SMEs can take restroom breaks whenever needed but he or she is required to have a card key to get back into the office.

The sessions usually begin with the examination facilitator training the SMEs about the background of criterion-referenced cutoff score measures. A detailed training outline is presented that explains the modified Angoff method. The outline encompasses the minimally acceptable and superior candidates, development of behavioral descriptions of each competency level, discussion on how to assign Angoff ratings, calibration session, rater monitoring, and instructions for rating each test item. SMEs are then asked to read a certain number of test items and write down their answers next to the test items. When every SME has completed the assigned numbers of test items, the correct answers to the test items are read aloud. SMEs are then asked to read the test items that they have just completed and then write down the probability that they think a minimally competent candidate would get the item correct. This probability is their initial rating for the test item. When the SMEs have completed their ratings for an assigned area, the test facilitator asks them to say aloud what their ratings are while he/she writes it down. Whenever there is a discrepancy between these initial ratings, a discussion will take place and the SMEs are then given the opportunity to revise their ratings so that the discrepancy can be resolved. A second rating is then collected from SMEs. The average of these final ratings is then used to set cutoff score for that test.
Data

This study utilized archival data from past Angoff passing scores workshops obtained from the Office of Professional Examination Services. The data consisted of the initial and the final ratings of SMEs, the initial mean and the final mean, and the probability values (p values) of previous candidates who got the test items correct. Numerical values were assigned to replace names of examinations and all identifiable information was kept anonymous and confidential and treated according to APA ethical standards.

Statistical Analysis

The study employed several statistical analyses. First, a one-way ANOVA was conducted to test for the reliability index of the data set. Next, SMEs’ initial ratings or pre-discussion ratings were analyzed using the descriptive statistics. This included their initial means, initial range, and their initial standard deviation. Assuming that group discussions followed SMEs’ initial ratings, the second task was to analyze SMEs’ final ratings or post-discussion ratings. This helped in identifying whether any revisions were made. This included SMEs’ final means, final range, and the final standard deviation. The next task was to analyze the initial and final means, the initial and final ranges, and the initial and final standard deviation to compare differences in SMEs pre-and post-discussions ratings. Separate columns were created for these differences and were called mean difference, range difference, and standard deviation difference.

Several paired sample t-test were used to identify significant changes in the pre- and-post means obtained. In order to test for the presence of conformity pressure and
group polarization effects, paired samples $t$-tests were carried out to explore changes in means and standard deviations of ratings. Mixed-model ANOVAs were also used to explore the degree to which these changes might differ across different examinations. A simple correlation analysis was also conducted to see the amount of changes that were made from pre-to-post-discussions. This included SMEs’ initial ratings and change scores (final ratings – initial ratings).

In order to test for the presence of minority pressure, another paired samples $t$-tests was conducted to explore changes in means. To test for the anchor and adjustment bias, descriptive statistics were conducted. If the means and standard deviations were the same from pre-to-post discussions, then SMEs ratings did not change. However, where there were changes in pre-and-post discussion means and standard deviations, this was interpreted as an indication of social influence during group discussions. Positive differences indicated that the ratings were revised downwards to lower ratings and negative differences indicated that ratings were revised upwards to higher ratings. The differences in means and standard deviation were used to test the hypothesized patterns. Finally, to examine if the non-independence assumption was violated, another paired samples $t$-tests was conducted.

Results

This study analyzed archival data from 38 Angoff passing score professional licensing examination workshops conducted by the Office of Professional Examination Services (OPES). The exams included a total of 3,983 test items or questions. Three test items were deleted because of an error in the data file. Ratings from SMEs ($N = 260$)
consisted of initial ratings ($M = 71.32$; $SD = 5.08$) and final ratings ($M = 71.25$; $SD = 4.88$). A reliability index developed by Shrout and Fleiss (1979) was used to assess the reliability of SMEs’ initial and final ratings. In order to do this, a one-way ANOVA was analyzed with raters as the grouping factor. The Shrout and Fleiss (1979) rater reliability index formula is as follows: $\{ICC \{1, K\} = (BMS – WMS) / BMS\}$, where $BMS$ is the between mean square and $WMS$ is the within mean square. The results showed a high reliability index of .96 for both initial and final ratings. Descriptive statistics summarizing the data set are provided in Table 1. The different procedures used in analyzing the data set, the results, and the interpretations are addressed separately for each hypothesis.
Table 1

*Descriptive Statistics for Initial Ratings, Final Ratings, and Item p-value.*

<table>
<thead>
<tr>
<th></th>
<th>Initial Ratings</th>
<th>Final Ratings</th>
<th>( p )-Value</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Mean</strong></td>
<td>71.32</td>
<td>71.25</td>
<td>68.58</td>
</tr>
<tr>
<td><strong>SD</strong></td>
<td>5.08</td>
<td>4.88</td>
<td>19.67</td>
</tr>
<tr>
<td><strong>Minimum</strong></td>
<td>53.33</td>
<td>54.00</td>
<td>0</td>
</tr>
<tr>
<td><strong>Maximum</strong></td>
<td>82.00</td>
<td>81.57</td>
<td>100</td>
</tr>
<tr>
<td><strong>Reliability</strong></td>
<td>.96</td>
<td>.96</td>
<td>-</td>
</tr>
</tbody>
</table>
Conformity Pressure

The conformity pressure hypothesis (hypothesis 1) predicted that following group discussions, SMEs with extreme initial ratings would revise their ratings inward toward the central tendency of the group, thereby suggesting the presence of conformity pressure. More specifically, it was predicted that if conformity pressure was the dominant social influence effect, the standard deviation would move inwards and a reduction in variance would be observed, while the means would remain the same. To examine these changes, descriptive statistics were computed that included SMEs’ initial means, initial range, and initial standard deviation. In order to identify whether any revisions were made, additional descriptive statistics were computed that included SMEs’ final means, final range, and final standard deviation.

To compare the differences in SMEs’ pre-and-post-discussions ratings, a paired sample t-test was computed. This test examined the mean difference between SME’s initial mean \( M = 71.34; \ SD = 7.39 \) and SMEs’ final mean \( M = 71.31; \ SD = 7.85 \). The mean difference between SMEs’ pre-and-post-discussions ratings was not significant \( t(3979) = .53, p = .595, \ d = .008 \). Cohen’s \( d \) indicates that there was a trivial effect size found between SMEs’ initial and final ratings. The Cohen’s \( d \) was calculated as follows: \( M / SD = 1.44 / 2.65 = .54 \), where \( M \) is the mean difference between SMEs’ initial and final ratings and \( SD \) is the standard deviation of the mean difference.

An additional paired sample t-test was analyzed to identify differences in variability as index by a change in the pre-to-post-discussions standard deviations. The test showed a significant reduction from SMEs’ initial standard deviation \( M = 7.87; \ SD \)
= 3.32) to their final standard deviations \((M = 6.43; SD = 2.04)\), \(t(3979) = 34.30, p = .000, d = .54\), thus indicating that SMEs with extreme initial ratings did revise their final ratings inward toward the central tendency of the group. Cohen’s \(d\) indicated that there was a medium effect size for change in standard deviations from pre-to-post-discussion.

The lack of changes in means and the reductions in standard deviations indicated that variability in SMEs’ final ratings shrank inwards after group discussions. This pattern in the data analysis supports the first hypothesis (conformity pressure). The presence of conformity pressure as indicated by the results supports the existence of social influence factors in Angoff passing score workshops.

In order to test the degree to which pre-to-post-discussion change varied from one SME panel to the next across examinations, a mixed ANOVA analysis was conducted. The between participants variable was the SME panel (exams), and the within participants variable was made up of the initial and final means. The summary of these results can be seen in Table 2.

The test of between subject effects indicated that there was no rater main effect \(F = (1, 37) = .71, p = .401, \text{partial } \eta^2 = .000\). The mean across initial and final ratings indicated that there were differences across cutoff scores between the examinations, \(F = (1, 37) = 42.42, p < .01, \text{partial } \eta^2 = .29\) with a 95% confidence interval ranging from .26 to .31 (See Table 2). This was an expected and important finding because harder exams should produce different cutoff scores than easier exams, and the grading should be harder on average. Furthermore, an interaction between ratings and exams was found, \(F = (1, 37) = 2.35, p < .01, \text{partial } \eta^2 = .02\) with a 95% confidence interval ranging from .00
to .02 (See Table 2). In other words, more change was observed for some examinations than others. In this study, 2% of the variability in change in ratings could be attributed to differences across exams \( (partial \eta^2 = .02) \), a small amount.
Table 2

*Analysis of variance for the degree to which pre-to-post discussion change varies across examinations.*

<table>
<thead>
<tr>
<th>Source of Variation</th>
<th>SS</th>
<th>df</th>
<th>MS</th>
<th>( F )</th>
<th>Partial ( \eta^2 )</th>
<th>95% C.I.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Exams</td>
<td>125,907.78</td>
<td>37</td>
<td>3,402.91</td>
<td>42.42*</td>
<td>.29</td>
<td>.26 - .31</td>
</tr>
<tr>
<td>Error(Exams)</td>
<td>316,212.64</td>
<td>3,942</td>
<td>80.22</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ratings</td>
<td>3.57</td>
<td>1</td>
<td>3.57</td>
<td>.706</td>
<td>.00</td>
<td></td>
</tr>
<tr>
<td>Ratings*Exams</td>
<td>438.16</td>
<td>37</td>
<td>11.84</td>
<td>2.35*</td>
<td>.02</td>
<td>.00 - .02</td>
</tr>
<tr>
<td>Error(Ratings)</td>
<td>19,892.79</td>
<td>3,942</td>
<td>5.05</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

*Note: \( N = 3,979 \)*

*\( p < .01 \)*
The result of these analyses can also be observed by looking at the plot of the difference cutoff scores in Figure 1 below. The results show that there were larger differences in pre-to-post-discussion means in some examinations than others; however, these differences were not large. The largest effect was the main effect of exams, meaning that the cutoff points for some examinations were higher than the cutoff points for other examinations.

Figure 1. Mean changes across examination cutoff scores.
Another mixed ANOVA was used to test the degree to which a reduction in standard deviation changes across examinations from pre-to-post discussion. The between participants variable was the SME panel (exams), and the within participants variable was made up of the initial and final standard deviations. The summary of these results can be seen in Table 3. The test of between subject effects indicated that there was a main effect found in the standard deviations across examinations, $F = (1, 37) = 1000.74, p < .01$, partial $\eta^2 = .20$, with a 95% confidence interval ranging from .17 to .22. The mean across initial and final standard deviations indicated that, averaged across all examinations, there were differences, $F = (1, 37) = 17.99, p < .01$, partial $\eta^2 = .14$ with a 95% confidence interval ranging from .11 to .15. Furthermore, an interaction between “standard deviation change” and exams was found, $F = (1, 37) = 13.55, p < .01$, partial $\eta^2 = .11$ with a 95% confidence interval ranging from .08 to .12. In this study, 11% of the variability in change in standard deviation could be attributed to differences across exams (partial $\eta^2 = .11$), a medium amount. The main effects of .11 gives a stronger interaction than that found between ratings and examinations.
Table 3

*Analysis of variance for the degree to which changes in standard deviation varies across examinations from pre-to-post discussion.*

<table>
<thead>
<tr>
<th>Source of Variation</th>
<th>SS</th>
<th>df</th>
<th>MS</th>
<th>F</th>
<th>Partial η²</th>
<th>95% C.I.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Exams</td>
<td>6,670.45</td>
<td>37</td>
<td>180.28</td>
<td>17.99*</td>
<td>.14</td>
<td>.11-.15</td>
</tr>
<tr>
<td>Error(Exams)</td>
<td>39,503.24</td>
<td>3,942</td>
<td>10.02</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SDchange</td>
<td>3,149.91</td>
<td>1</td>
<td>3,149.91</td>
<td>1000.74*</td>
<td>.20</td>
<td>.17-.22</td>
</tr>
<tr>
<td>SDchange*Exams</td>
<td>1,578.45</td>
<td>37</td>
<td>42.66</td>
<td>13.55*</td>
<td>.11</td>
<td>.08-.12</td>
</tr>
<tr>
<td>Error(SDchange)</td>
<td>12,407.82</td>
<td>3,942</td>
<td>3.15</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

*Note: N = 3,979*

*p < .01*
The result of these analyses can also be observed by looking at the plot of the different standard deviation of the initial and final ratings across cutoff scores in Figure 2 below. The results show that the initial ratings, as seen by the solid lines, had a higher standard deviation than the final ratings, as seen by the dashed lines. This observation reveals consistent changes of the reduction in standard deviation in final ratings across examinations as compared to the initial ratings. However, there was also a significant interaction found between SMEs’ ratings and standard deviation changes across all examinations, meaning that the amount of initial-to-final change differs across examinations.

**Figure 2.** Standard Deviation changes from initial (1) to final (2) ratings across examination cutoff scores.
Group Polarization

The group polarization hypothesis predicted that there would be evidence that SMEs’ mean ratings from pre-to-post discussion would become more extreme towards the direction that the group tended and variability among group members would at the same time decrease. This hypothesis was tested by analyzing SMEs initial and final means, and initial and final standard deviations. The test examined the mean difference between SME’s initial mean \((M = 71.34; SD = 7.39)\) and SMEs’ final mean \((M = 71.31; SD = 7.85)\). As found in the previous hypothesis, there were no differences in the means but there was a reduction in the standard deviation. From the result of the conformity pressure paradigm, the mean of the initial standard deviations was 7.87 and the mean of the final standard deviations was 6.43. The results of both the \(t\)-test and mixed ANOVA supported the presence of conformity pressure, and refutes the conclusion that group polarization was present in Angoff passing score workshops.

To further examine this hypothesis from a different perspective, a simple correlation analysis was conducted between “a change score” (i.e. final ratings – initial ratings) and the initial ratings for each test item. As seen in Figure 3, a negative correlation \(r(26537) = -.37\) was obtained, indicating that ratings that were initially relatively higher tended to be changed to lower ratings and ratings that were initially lower, tended to be changed to higher ratings. This finding was again more indicative of the conformity pressure effect than the group polarization effect.
Figure 3. A correlation between initial ratings and change score.

Figure 4 shows a histogram of the changes made in ratings from pre-to-post-discussion. When looking at the histogram, the data show that although most ratings did not change from pre-to-post-discussion, the few amount of changes made tended to fall within the range of approximately negative 30 points to positive 30 points (-30 to +30) of the data from pre-to-post-discussion. These changes can be observed on either side of the graph. On average, extreme initial ratings at either end of the scale tended to be revised inwards towards the middle.
Figure 4. Differences in ratings from pre-to-post-discussion
Minority Pressure

The minority influence paradigm predicts that an individual group member can successfully persuade the majority members of the group to accept his or her minority view point. An observable shift in the final ratings of group members towards that of the minority opinion would provide evidence of a single SME influencing others during group discussions.

For the purposes of this study, the minority opinion holder was defined as the SME whose initial ratings deviated the most from others in the group. Minority opinion holders were identified in the data set by computing the mean of the rest of the group with the minority person removed from the initial group mean. Next, each SME’s deviation from the minority was computed on both the initial and final ratings. By doing this, a new data set was created that showed the mean deviation of other group members from the minority in both initial and final ratings. Finally, a paired sample $t$-test was conducted to determine whether there was a significant change in the final ratings towards the minority.

The paired sample $t$-test examined the mean difference between SME’s initial deviation from the minority ($M = 16.02; SD = 6.76$) and SMEs’ final deviation from the minority ($M = 15.45; SD = 6.61$). The mean difference between SMEs’ pre-and-post-discussion deviations from the minority was significant $t(3109) = 10.42, p < 0.000, d = .19$. This implies that SMEs final ratings moved closer to that of the minority opinion holder after group discussions. Cohen’s $d$ indicated that there was a small effect size found between SMEs’ initial and final deviation from the minority. However, as will be
discussed later, it is difficult to separate a shift toward the minority from a shift toward the group mean.

*The Insufficient Anchor and Adjustment Bias*

The insufficient anchor and adjustment bias predicted that during estimation, SMEs would retrieve information that they know from memory and utilize that information as an anchor for their initial estimates. It was hypothesized that there would be a consistent mean across raters and the judgments SMEs provide would not deviate much from that mean, suggesting an insufficient anchor and adjustment bias. This hypothesis was tested by computing the means for each SMEs’ initial ratings across all items they rated, and then summarizing those means with descriptive statistics ($M = 71.34; SD = 7.39$); SMEs’ final ratings ($M = 71.31; SD = 7.85$) and comparing them to similar statistics for the actual probability estimates ($p$-values) from previous examinations ($M = 68.58; SD = 19.67$). The results indicated that the means across raters were very consistent and were near 70% as predicted and the standard deviations did not deviate much from pre-to-post-discussion. Compared to the actual item difficulty level, higher differences in variability were found between the judgments provided by SMEs and the actual level of difficulty of the test items. Differences found in initial ratings were 7.08 times higher and differences found in final rating were 6.28 times higher than those from previous examinations. These differences were obtained by computing the $SD^2_{p\text{-value}} / SD^2\text{MeanInitial}$, and $SD^2_{p\text{-value}} / SD^2\text{MeanFinal}$. This result supports the insufficient anchor and adjustment bias.
Addressing the Non Independence Assumption

Stevens (2007) contended that in situations where a data set is comprised of responses from several small groups, a correlation is often observed within groups. Since the correlation does not occur across groups, he recommended that the group’s mean be used as the unit of analysis to test for the presence of non-independence between observations. Because this study analyzed examination data of small groups, non-independence issues may have potentially affected the results creating the possibility of an inflated type 1 error. Two \( t \)-tests were performed to see if the pattern of the results were the same as their corresponding \( t \)-tests computed earlier at the level of individual ratings. A different pattern would indicate that the non-independence assumption was violated. In order to do this, the data were collapsed down to the group level so that each group was analyzed only once. This was important because at the examination level, the \( t \)-test included all the test items and ignored that some of the test items were in the same group.

To compare the differences in SMEs’ pre-and-post-discussions ratings, a paired sample \( t \)-test was computed. This test examined the mean difference between SME’s initial mean (\( M = 71.34; SD = 4.41 \)) and SMEs’ final mean (\( M = 71.31; SD = 4.41 \)). The mean difference between SMEs’ pre-and-post-discussions ratings was not significant \( t(38) = .68, p = .50, d = .11 \). Cohen’s \( d \) indicates that there was a trivial effect size found between SMEs’ initial and final ratings.

An additional paired sample \( t \)-test was analyzed to identify differences in variability in the pre-to-post-discussions standard deviations. The test showed a
significant reduction from SME’s initial standard deviation ($M = 8.06; SD = 1.40$) to their final standard deviations ($M = 6.55; SD = .78$), $t(38) = 10.38, p = .000, d = 1.68$. Cohen’s $d$ indicated that there was a large effect size for change in standard deviations from pre-to-post-discussion. The results of these $t$-tests indicate that the same effect occurred at the group level when the data were aggregated across individual observations.

General Discussion

This study investigated evidence of social influence factors and cognitive bias in the Angoff method of setting passing scores through the analysis of pre-and post-discussion rating data. More specifically, this study investigated how social influence factors and cognitive bias might influence the decisions of SMEs in Angoff passing score workshops. Conformity pressure, group polarization, and minority pressure were the social influence factors that were investigated. The insufficient anchor and adjustment bias was the cognitive bias that was investigated. This study has reviewed three important social influence factors and a cognitive bias that might likely occur in an Angoff passing score workshop. As part of the main discussion, a distinction was made between cognitive biases and social influence factors.

Hypothesis 1 predicted that following group discussions, SMEs with extreme initial ratings would revise their ratings inward toward the central tendency of the group, thereby suggesting the presence of conformity pressure. More specifically, it was predicted that the standard deviation would move inwards and a reduction in variance would be observed, while the means would remain the same. The results from hypothesis 1 revealed that SMEs with extreme ratings in Angoff passing score workshops did tend to
revise their final ratings inwards towards the central tendency of the measuring scale.
There were no significant differences found between the means of the initial ratings and that of the final ratings. However, significant differences were found between the initial standard deviation and the final standard deviation. The results showed that SMEs revised their final ratings inwards towards the central tendency of the group, thus, indicating the presence of conformity pressure during group discussions. The findings in this study supported the conformity pressure hypothesis.

Hurtz and Auerbach (2003) asserted that group discussions in Angoff passing score setting are beneficial if the discussions involve exchange of information between experts. Furthermore, other studies have found that group discussions help in reducing inter-judge variability in judges (Hambleton & Plake, 1995; Hurtz & Auerbach, 2003). However, the reduction in variance found in this study may have been due to OPES policies. In particular, the OPES - TVDS staff typically employ a “range policy” that requires SME ratings to fall within a specific range before moving on to the next item. While some staff maintained a range policy of 20%, others maintained a range policy of 25%. Whenever SMEs initial ratings did not fall in the desired range, group discussions were held and SMEs were given a second chance to revise their ratings. Typically, in workshops, SMEs with extreme ratings are asked to explain the rationale behind their ratings, thereby putting them “on the spot”. Oftentimes, SMEs in this situation would revise their ratings to conform to that of the group. This could be seen in the differences in the initial range and the final range. Initial ranges with higher numbers tended to be shrunk in the final ranges.
Maintaining a range policy has the advantages of making group consensus more easily achieved, and therefore speeding the process of determining a cutoff score. However, asking SMEs to reach a range before moving on to the next test item may have forced them to change their ratings so as to adhere to a consensus. By imposing “a conformity pressure” policy to get SMEs to reach a consensus might also undermine the expertise of SMEs. Although the job of the SME is to set standards, the tension between the desire to set these standards and the social pressure to be viewed by others as doing a good job may interfere with the judgments that they make, and therefore undermine the desired cutoff score.

It would be interesting to compare the results of this study with Angoff workshop data from other organizations that did not maintained this range policy. Future studies should replicate this study using Angoff passing score data that did not maintain this range policy.

Group polarization hypothesis predicted that there would be evidence that SMEs’ mean ratings from pre-to-post discussion would become more extreme towards the direction that the group tended to agree to, and variability among group members would at the same time decrease. The group polarization effect contends that SMEs who had rated an item as moderately difficult in their initial ratings, may revise their ratings after group discussions to be more extreme towards the direction that the group tended to agree to. This hypothesis was tested by analyzing SMEs initial and final means. There were no significant changes found in the means. However, there were reduction in the standard deviations of data that was more indicative of the conformity pressure paradigm, thus,
refuting the presence of group polarization effect. A simple correlational analysis also
did not show any evidence of group polarization effects in Angoff passing score
workshops.

The lack of evidence for the group polarization effects was inconsistent with
Myers and Lamm (1976) description of the risky shift phenomenon. This phenomenon
assumes that a group’s overall tendencies will become more risk-seeking than the initial
average risk seeking tendencies of the individual members. However, the findings of this
study may have been influenced by the range policy discussed earlier. Moreover, it is
possible that the data used in this study may not be the appropriate type of data set for
observing this type of social influence factor again because of the range policy that is
designed for groups to reach consensus. Future research should utilize passing score data
from other types of criterion-referenced testing such as the Bookmark method to see if
different results are obtained.

The minority opinion hypothesis predicted that an individual group member with
strongly held views would be able to successfully persuade the majority members of the
group to accept his or her viewpoint during group discussion. The results of the paired
sample t-test were significant, implying that after group discussions, there was a tendency
for SMEs to revise their final ratings towards that of the minority. Gordijn, de Vries, and
de Dreu (2002) contended that minority influence pressures are strongest in cases where
there is consistency in the minority message or when the minority position is different
from that of the norm. Although the data analysis of this study provided support for the
minority opinion hypothesis, it is impossible to say so conclusively because the move
toward the minority may also reflect a movement toward the group central tendency (since the group mean will fall between the majority and the minority outlier), as was the case with the conformity pressure paradigm. This could be observed indirectly in the differences in effect sizes obtained. The $d$-statistic for the reduction in deviations from the mean (conformity pressure) was larger than the $d$-statistic for the reduction in deviation from the minority (minority influence), perhaps suggesting that the minority influence effect here was confounded with a simple inward shift. In his experimental study, Pasisz (2006) also found that judges were more likely to change their judgments towards the mode of the group. Future research should be conducted to see if similar results would be obtained with passing score data that does not require a movement toward the group central tendency.

The probability that an SME will adjust his or her final ratings toward the minority’s may be influenced by the experience of the minority viewpoint holder. The arguments of a highly experienced SME are likely to be more persuasive in group discussions than those of less experienced SMEs. For example, an SME who is consistent with his/her message may say, “I do this everyday at work, and I have been on the field for 20 years now”. A minority opinion holder’s length of experience and consistency in message may successfully convince others in the group to accept his or her point of view. Unfortunately, it was impossible to test this in the data that were available for this study.

Future research into the effects of minority pressure in Angoff passing score workshops should test whether there is a correlation between the personal attributes of a
minority opinion holder and the tendency of a group to move their ratings towards the minority’s. The work experience of the minority rater would be a key variable of interest.

This study also examined what Tversky and Kahneman (1974) called the anchoring and adjustment heuristic. The estimates of SMEs in several Angoff passing score workshops were tested for evidence of insufficient anchoring and adjustment bias. The insufficient anchor and adjustment bias predicted that during estimation, SMEs would retrieve information that they know from memory and utilize that information as an anchor for an initial estimate. This hypothesis predicted that there would be a consistent mean across raters and that the judgments of SMEs would not deviate much from that mean.

The findings in this study provided support for this hypothesis. Initial and final means were consistent and the standard deviations did not deviate much from pre-to-post-discussions. However, the standard deviation of the $p$-values of the test items was much higher than that for SMEs’ ratings of item difficulty, thereby revealing a discrepancy between SMEs’ judgments and the actual range of difficulty levels of the test items. This supports the presence of the insufficient anchor and adjustment bias.

The findings of this study showed evidence that SMEs insufficiently adjusted their final estimates from that of their initial estimates. This was probably because when SMEs found that their estimates were within the desired range, they stopped adjusting their estimates. This finding was consistent with research in this arena (see Epley &
Gilovich, 2001, 2006; Tversky & Kahneman, 1974). This finding provided support for the presence of cognitive biases that may be present in Angoff passing score workshop.

These findings suggest that the insufficient anchor and adjustment bias may be present in Angoff passing score workshops. It is possible that human biases can influence the development of task-level estimates and these biases can negatively impact the estimates obtained. Although it is impossible to eliminate these biases, awareness, understanding, and the incorporation of bias-reduction strategies such as proper training can help mitigate these negative impacts.

Further study will be required to determine to what extent the insufficient anchor and adjustment bias may impact SMEs ratings and the overall cutoff score acquired. Because of the importance attached to standard setting, it is very important that SMEs in passing score workshops receive thorough training on how to assign ratings for the test items. Maurer, Alexander, Callahan, Bailey, and Dambrot (1991) suggested that thorough training should be given to SMEs before they begin the process of determining a cutoff score.

During the training session, greater emphasis should be given on the distinction between the highly competent and the highly incompetent. However, the focus should be on the minimally competent candidate. Frequent reminders should also be given throughout the workshops that remind SMEs that their focus should be on how the barely / minimally competent candidate would perform on a particular test item, and not on what candidates should know.
Regarding the anchor and adjustment bias, there was a discrepancy found in the variability between the judgments provided by SMEs and the actual level of difficulty ($p$-values) of the test items. Maybe showing SMEs the $p$-values from previous exams might help them in assigning their ratings in ways that more accurately reflect the difficulty level of the test items.

It is possible that utilizing data from several different workshop facilitators and several different types of licensing exams may have affected the results obtained in this study. Future research should use data from single exam facilitator to see if different results are obtained.

It is important to stress that most research in cognitive biases have adopted the traditional cognitive perspective with social cognitive factors largely being neglected. It is also possible that some biases have more potential to affect group decision making than others. Depending on the purpose of the group, cognitive biases and social influence factors may affect individuals differently and to varying degrees. Time pressures and the importance of achieving a consensus within groups are factors that may be stressful in a passing score workshop. Exam facilitators should do their best to make SMEs feel relaxed during standard setting workshops.

Finally, it is also important to understand how situational factors relevant to group decision making may interact with cognitive and social influence factors in affecting the cutoff score acquired. More research should also be conducted that investigates what other factors influenced SMEs judgments’ in determining Angoff cutoff scores, such as background factors that may interfere with SMEs’ judgments.
REFERENCES


