RE-CONCEPTUALIZATION OF MODIFIED ANGOFF STANDARD SETTING: UNIFIED STATISTICAL, MEASUREMENT, COGNITIVE, AND SOCIAL PSYCHOLOGICAL THEORIES

By

Ifeoma Chika Iyioke

A DISSERTATION

Submitted to
Michigan State University
in partial fulfillment of the requirements
for the degree of

Measurement and Quantitative Methods - Doctor of Philosophy

2013

ABSTRACT

RE-CONCEPTUALIZATION OF MODIFIED ANGOFF STANDARD SETTING: UNIFIED STATISTICAL, MEASUREMENT, COGNITIVE, AND SOCIAL PSYCHOLOGICAL THEORIES

By

Ifeoma Chika Iyioke

This dissertation describes a design for training, in accordance with probability judgment heuristics principles, for the Angoff standard setting method. The new training with instruction, practice, and feedback tailored to the probability judgment heuristics principles was called the Heuristic training and the prevailing Angoff method training was called the Normative training.

To evaluate effectiveness of the Heuristic training over the Normative training, the researcher ran two empirical studies for this dissertation. The design of the empirical study was a two-way mixed factorial effect ANOVA (2 training methods × 3 rounds of judgment). The empirical studies recommended cut score for the fourth grade mathematics, Michigan Educational Assessment Program (MEAP), Proficient performance category.

There were 10 and 12 participants in the Heuristic training and the Normative training, respectively. The participants of the studies were comprised mostly of Michigan State University (MSU) pre-service teachers and teachers in the mid-Michigan area. Two tests were used for the judgments, one for the practice round of judgment called the Practice test and the other for the feedback rounds of judgment called the Real test. Both the Practice and the Real test were comprised of subsets of released Michigan Educational Assessment Program (MEAP), 2005-2006 fourth grade multiple choice items. There were three rounds of probability judgments in both studies namely: the practice and the two feedback rounds. For both training methods, the practice round encompassed pre-requisite tasks for probability judgment and ensued instruction while the feedback rounds followed feedback to the participants on their judgments in the preceding round. The Heuristic and the Normative training methods were evaluated for substantive meaningfulness of the probability judgments and their cut scores derivatives, in relation to the heuristic model assumptions. The methods were also compared for the effectiveness of training interventions of instruction and practice versus feedback.

In the practice round of judgment, the training groups performed comparably, the participant's judgments fit the probability judgment heuristic principles, and cut scores were quite reasonable for both groups. Conversely, in the feedback rounds, the participant's judgments deviated from the probability judgment heuristic principles, were considerably less substantively meaningful for the Normative training, and cut scores were positively biased for both groups.

The conclusion based on overall findings were that the Heuristic training was more effective than the Normative training and that regardless of training method that instruction and practice activities were more effective than feedback. Intellectual merits of the dissertation, recommendations, study limitations, and directions for future Heuristic training Angoff studies are discussed.

DEDICATION

This dissertation is dedicated to God Almighty. His banner over me is love. It is through Him that I can achieve excellence through honest labor.

ACKNOWLEDGEMENTS

My love and affection is expressed to my husband, Ike Iyioke, and my sons, Ikenna Iyioke, Arinze Iyioke, and Ililochi Iyioke for their moral support and patience with me throughout these years of hard work. I express my profound appreciation of my parents Dr. John Ibemesi and Dr. Felicia Ibemesi, both also alumni of MSU, whose inspiration apparently paved the way for me.

Also, my gratitude goes to all my committee members for their shared insights to make this work possible: my dissertation director, Dr. Mark Reckase, whose teaching highlighted gaps in the standard setting literature that inspired me to take on the topic; and, for accepting to take up the role of dissertation director at the abrupt departure of my former advisor. Next, I am very appreciative of my former advisor Dr. Sharif Shakrani, for accepting me as his student thus making it possible to pursue my doctoral degree; Dr. Edward Roeber, whose invaluable support and timely guidance made it possible to navigate the logistics of empirical research; Dr. Devin McAuley, who crystalized research paradigms in psychology, and Dr. Tapabrata Maiti, whose statistical input provided insight on the subject matter.

A lot of other people at MSU also made this work possible including, Dr. Todd Drummond, Lihong Yang, Steven Viger, Dr. Adam Wyse, Dr. Richard Houang, Kathleen Whyte, and Dr. Timothy Pleskac. Likewise, I am very grateful to Dr. Thomas Hirsch, Dr. Robert Keene, Dr. Micheal Bunch, and Dr. Micheal Beck, all standard setting practitioners, for sharing their insights with me. There are many people too numerous to mention that helped me along the way in my graduate studies, to all of you, I say thank you.

PREFACE

This dissertation represents an attempt to synthesize theoretical frameworks relevant to the Angoff standard setting method. In addition, it demonstrates an approach to design a training program based on the theoretical frameworks and for evaluating its effectiveness.

The import of this dissertation resides in the depth and breathe of topics covered and particularly formulated in sufficient details than would be possible with a journal article. As such, it is meant to be a useful compendium of ideas to inform future standard setting research and practice. Meanwhile, a quick preview of the content follows.

The text of the dissertation is comprised of eight chapters. Chapter one introduces the topic of standard setting. Chapter two accounts for the measurement, statistical, and, psychological theoretical frameworks for the Angoff standard setting method. Chapter three discusses research on the Angoff method. Chapter four articulates re-conceptualizations of the Angoff method based on the preferred theoretical positions. Chapter five considers the evaluation and analytic frameworks for training. Chapter six delineates the methods of the dissertation's empirical studies. Chapter seven presents the results of empirical studies. In Chapter eight, intellectual merits of dissertation, findings, interpretation of the findings, recommendations, limitations and directions for future research and practice are discussed. The reference materials include definitions of major concepts, scripts and instruments used for the empirical studies and a rich bibliography.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

KEY TO ABBREVIATIONS

AN	Data Interpretation and Analysis
AYP	Adequate Yearly Progress
CTA	Cognitive Task Analysis
D	Data and Probability
DOK	Depth of Knowledge
ET	Extended Thinking
FL	Fluency With Operations and Estimation
G	Geometry
GLCE	Grade Level Content Expectation
GS	Geometric Shape, Properties, and Mathematical Arguments
IRB	Institutional Review Board
IRT	Item Response Theory
KSA	Knowledge, Skills, and Abilities
LO	Location and Spatial Relationships
M	Measurement

MCF	Michigan Curriculum Framework
MDE	Michigan Department of Education
ME	Meaning, Notation, Place value, and Comparisons
MEAP	Michigan Educational Assessment Program
MR	Number Relationships and Meaning of Operations
MRM	Multifaceted Rasch Model
MSU	Michigan State University
N	Number and Operations
NAEP	National Assessment of Educational Progress
NCLB	No Child Left Behind
NSF	National Science Foundation
NYSCI	New York Hall of Science
OIB	Ordered Item Booklet
PCOA	Principal Coordinates Analysis
PLD	Performance Level Descriptor
PR	Probability
PS	Problem Solving Involving Measurement

RE	Recall
RE	Data Representation
REESE	Research and Evaluation on Education in Science and Engineering
RP	Response Probability
SC	Skills and Concepts
SPSS	Statistical Package for Social Sciences
SR	Spatial Reasoning and Geometric Modeling
ST	Strategic Thinking
TE	Techniques and Formulas for Measurement
TR	Transformation and Symmetry
UN	Units and Systems of Measurement

Chapter One: Introduction

Cizek (2001) defined standard setting as the task of determining the levels of performance on educational or professional tests that allows inferences about students such as, classifying them to performance categories based on their demonstrated knowledge and skills competencies. In the United States licensure and certification educational settings, where interest usually is in making pass or fail decisions, which scenario presents the simplest case of standard setting, the derived levels of performance or minimum test score for making the classification decision is referred to as the passing score. In the public school contexts, where increasingly standard setting for multiple performance categories is becoming the norm rather than the exception, these derived levels of performance or minimum test scores established between adjacent performance categories (with the exception of the lowest performance category), are called the cut scores. Standard setting is an essential aspect of educational measurement because its proceeds figure in making decisions in diverse educational settings that includes the aforementioned licensure, certification, and public school contexts. Because of the role of standard setting in educational decision making, it is important that the process generates accurate and reliable results.

There are two approaches to standard setting namely: norm-referenced and criterion-referenced approaches. The norm-referenced approach to standard setting implicates *a posteriori* determination of the passing or cut scores of performance categories relative to or dependent upon the performance of a group of students on a test. The criterion-referenced approach to standard setting involves *a priori* determination of standards of performance usually specified in terms of knowledge and skills to be measured so that the passing or cut scores are operational versions of these standards of performance.

The norm-referenced approaches to standard setting are still very much in use in low-stakes examination settings such as in formal educational training settings that include K-12 classrooms and tertiary institutions. Conversely, the norm-referenced approaches to standard setting are gradually been phased out in high-stakes examination contexts that include the licensure, the certification, and in the public school accountability standard setting contexts (e.g. *NCLB* school accountability initiatives). Interest in the licensure, the certification, and in the public school accountability examination contexts have changed to measuring examinees competence with respect to the pre-determined standards of performance (Cizek, 1991). As a consequence, the criterion-referenced standard setting approaches are burgeoning in these high-stakes standard setting contexts and are typically performed in laboratory settings where a group of participants are called together for the purpose of deriving the performance levels.

According to Cizek (1991), the primary goal of standard setting that empanels participants for laboratory standard setting is variance reduction on the final judgment outcomes. This goal of variance reduction is accomplished through the introduction of group process procedures. Cizek (1991) however, identified the problems arising from these group process procedures. The highlighted problems include the tedium of the tasks plus the consensus reaching procedures, the cost involved, and the potential negative effect of group process procedures on the resulting cut scores. In spite of these practical challenges, this state of laboratory standard setting practice affairs persists. Consequently, to be researched in this dissertation is the criterion-referenced laboratory standard setting and for the high-stakes public school accountability context, although referred to for short as standard setting. However, before foraying into formal description of laboratory standard setting research concepts and theories, it is appropriate to begin with the statement of goal of this dissertation.

To lay the groundwork for the goal of this dissertation was the analogy between the problem of standard setting and that of legal practitioners of deciding on where and how to draw the line (Camilli, Cizek, & Lugg, 2001; Cizek, 1993; Cizek, 2001; Cizek & Bunch, 2007; Lerner, 1979). Given this analogy between standard setting and legal process, it is important to highlight that the offshoot of field approaches to study of eye witness memory in the cognitive psychology field was the goal of veridicality. According to Eysenck and Keane (2010), citing Koriat and Goldsmith (1996), while traditional laboratory approach to study of eyewitness memory was based on the store house metaphor, what matters is simply how many items of information can be recalled. Everyday memory field research was based on the correspondence metaphor according to which, the content of what is remembered was more important.

Change of goal and course in researching eye witness memory mattered because innocent people have been imprisoned solely on the basis of eyewitness testimony. Likewise, a change of goal and course in researching laboratory standard setting judgment process is necessitated because standard setting outcomes figure in decision making in a variety of educational settings. The proceeds of standard setting are used in these settings in making high-stakes decisions such as, in selecting who receives an educational intervention or who practices a profession, that impacts the lives of the individuals involved in meaningful ways (Cizek, 2001; Zieky, 2001).

Therefore, this dissertation prescribes actions for the high-stakes public school standard setting contexts that are predicated upon the assumption that participants are knowledgeable about empirical facts. The primary goal of prescribing these set of actions is to enhance veridicality of laboratory standard setting research outcomes, while the secondary goal is to increase efficiency in the implementation of the standard setting process. To begin to formalize this dissertation research, discussed in the section that immediately follows are concepts and processes of laboratory standard setting research.

1.1. Laboratory Standard Setting Research Paradigm: Concepts and Processes

This section serves to review current laboratory standard setting research concepts and processes.

1.1.1. Performance Standards

Reckase (2001), citing the Webster's New Collegiate Dictionary (1977), defines a "standard" as "something set up and established by authority as a rule for the measure of quantity, weight, extent, value, or quality" (p. 1133). This concept of standard was adapted to the United States K-12 high-stakes public school educational testing context. Policy makers put accountability systems (e.g. *the No Child Left behind Act*) in place that set goals for school organization improvement by delineating standards of performance (Fuhrman & Elmore, 2004; Jacob, 2005; Moe, 2003; O'Day, 2002; Peterson & West, 2003; Rudalevige, 2003). The performance standard in the high-stakes public school testing contexts refers to qualitative descriptions of intended distinctions between adjacent categories of students' performance (Kane, 2001). They are articulated in the form of policy by the agency that calls for the existence of the standard, with the term agency synonymous with authority referred to by the dictionary definition (Reckase, 2001; Reckase, 2009). The agencies in the case of public school NCLB accountability program are the state departments of education under the auspices of the U.S. department of education (Perie, 2008).

From the measurement perspective, the performance standard refers to the intended result of the policy making agency calling for the standard, while standard setting methods are means for accomplishing the result (Reckase, 2001). For the purpose of measurement, performance standards are delineated in performance level descriptors (PLDs). The "PLDs describe the amount of knowledge and skills required of each performance level" (Perie, 2008, p. 15). In the public school contexts, knowledge and skills of a content domain are articulated in a curriculum framework prior to the establishment of PLDs. Consequently, PLDs are delineated for each performance level and grade level in terms of the knowledge and skills addressed by the content domain of the subject area (e.g. mathematics). To instantiate a PLD, the 2005 Michigan Educational Assessment Program (MEAP) mathematics PLD used in this dissertation comprised of four PLDs namely: Exceeded Michigan Standard, Met Michigan Standards (i.e., the Proficient performance level), Basic, and Apprentice. Although performance standard and PLD's refer to the same thing and could be used interchangeably, for the purpose of presentation of the conceptual pieces in this introductory section, the performance standards would be used while PLD is reserved for the method section.

Meanwhile, it is assumed for the purpose of measurement that performance standards reflect the quantities of knowledge and skills of performance levels intended by the agency. The fundamental research problem therefore for educational measurement is how to measure the quantities presumably underlying the performance standards. Researchers in educational measurement research field devise approaches to solving this problem and these approaches are called standard setting methods, however referred to for short as standard setting. The immediately following section reviews conceptual views about standard setting.

1.1.2. Standard Setting

There are two conceptual views about standard setting, namely: parameter estimation and value judgment. This section briefly reviews these two conceptual views of standard setting in their order of historical ascendancy.

The nascent view of standard setting was as a process that parallels estimation of a population parameter so that it was believed that there is a theoretically correct value for the cut score, just as there is a true value for any statistic in the population (Cizek & Bunch, 2007; Jaeger, 1989; Jaeger, 1991; Zieky, 2001). The contemporary view of standard setting is of a process that evokes value judgment, so that it is believed that cut scores are constructed, not found, and that a right answer does not exist (Cizek, 2001; Cizek & Bunch, 2007; Zieky, 2001).

As you will see later in the next chapter on the theoretical framework for the Angoff (1971) standard setting method, while the parameter estimation view of standard setting is consistent with the realist measurement and the objective probability research paradigms, in contrast, value judgment is consistent with the operational measurement and the subjective probability paradigms. Meanwhile, the conceptual view consistent with the goal of this dissertation research of enhancing veridicality of laboratory standard setting outcomes is parameter estimation. Henceforth, all subsequent propositions about standard setting in this dissertation are based on the perspective of parameter estimation. As a consequence, Reckase's (2009) parameter estimation model for standard setting would be the conceptual reference for the rest of the discussion in this introductory section. Before proceeding to explicate Reckase's parameter estimation model for standard setting, it is important to highlight two more important conceptualizations of laboratory standard setting methods:

First, standard setting methods are classified based on the object focus of judgment as examinee-centered or test-centered. The term “object” is used to refer to examinees and test items. For examinee-centered methods the judgment focus is on content domain knowledge and skill attribute of the examinees in relation to the performance standard. Likewise, the judgment focus for test-centered methods is on content domain knowledge and skill attribute of test items in relation to the performance standard. With regards to this conceptualization of standard setting methods, it is important to highlight that the focus of this dissertation is the test-centered methods. Precisely, the Angoff method, is an example of a test-centered method (Cizek, 2001; Cizek & Bunch, 2007; Hambleton & Pitoniak, 2006; Jaeger, 1989; Wyse, 2009). Besides, the Bookmark method, another method that would also be made reference to in this dissertation, is equally a test-centered method (Mitzel, Lewis, Patz, & Green, 2001).

Second, standard setting was conceptualized as a stimulus-centered measurement process. As such, standard setting is analogous to student performance assessment. The construct measured in standard setting is participant’s mental representation of student performance at the threshold of achievement level (Nichols, Twing, Mueller, & O’Malley, 2010). The concept of stimulus refers to all materials presented to standard setting participants for which measures are derived. In the case of test-centered methods the stimuli includes PLD and test. The concept of stimulus is more encompassing and is consistent with the cognitive psychology theoretical framework of this dissertation. Therefore, test-centered standard setting methods that include the Angoff method would be regarded as stimulus-centered methods.

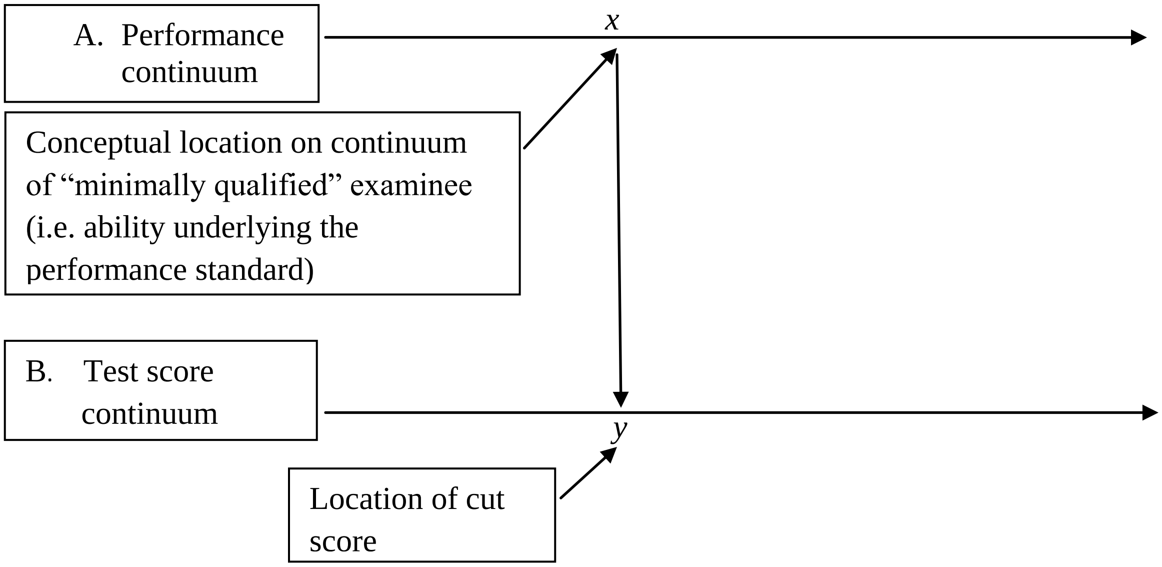
1.1.3. Parameter Estimation Model for Standard Setting (Reckase, 2009)

Because of the Angoff method focus of this dissertation, all discussions about the Reckase’s (2009) parameter estimation model for standard setting in this section pertain to and generalize directly to the stimulus-centered methods. Reckase’s (2009) parameter estimation model for standard setting describes standard setting methods as an organized system for collecting the judgments of qualified individuals for translating from the language of policy to that of test score scale (Reckase, 2001; Reckase, 2009). The model assumes that the agency that calls for a performance standard has some intended ability level in mind when they articulated the policy so that they could tell when the outcome of standard setting does not match their intention (Reckase, 2009). Hence, the assumption is that underlying each performance standard is a representative and quantifiable ability of the performance category described. Conceived thus, standard setting process parallels population parameter estimation (Zieky, 2001). Moreover, according to the model, standard setting is more appropriately called standard translation. Therefore, the model also assumes that the performance standard is in place before standard setting and that tests are developed for the purpose of standard setting that samples the knowledge and skills of the performance standards.

Consequently, standard translation involves translating the language of the policy to that of numerical test score and was used in a metaphorical sense to draw the analogy between the standard setting process and that of translating text from one language to another (Reckase, 2009). The individuals who translate standards are called panelists or judges. However for sake of generality, these individuals would be referred to as “participants” in this dissertation. The judgments of these qualified individuals are typically guided by the performance standard and test which are the stimuli. By virtue of involvement of these individuals in the standard translation process implies that standard setting is psychological measurement. The usual product of the standard translation process is the cut score which is a score on the reporting scale for the test that presumably represents the performance standard. In this sense, standard translation establishes the logical relation between the presumed intended ability quantity underlying the performance standard and the cut score established on a test. This logical relation, between a performance standard and cut score is represented pictorially in the Figure 1-1 that immediately follows.

Figure 1-1 illustrates the relationship between performance standard and cut score as adapted from Kane (1994). Panel A shows the performance continuum while panel B shows the test score scale. This illustration postulates that participants conceptualize the point *x* on the performance continuum which separates those that meet the performance standard for the testing purpose from those that do not and subsequently, they represent this conceptualization in terms of a cut score *y* on the test score scale.

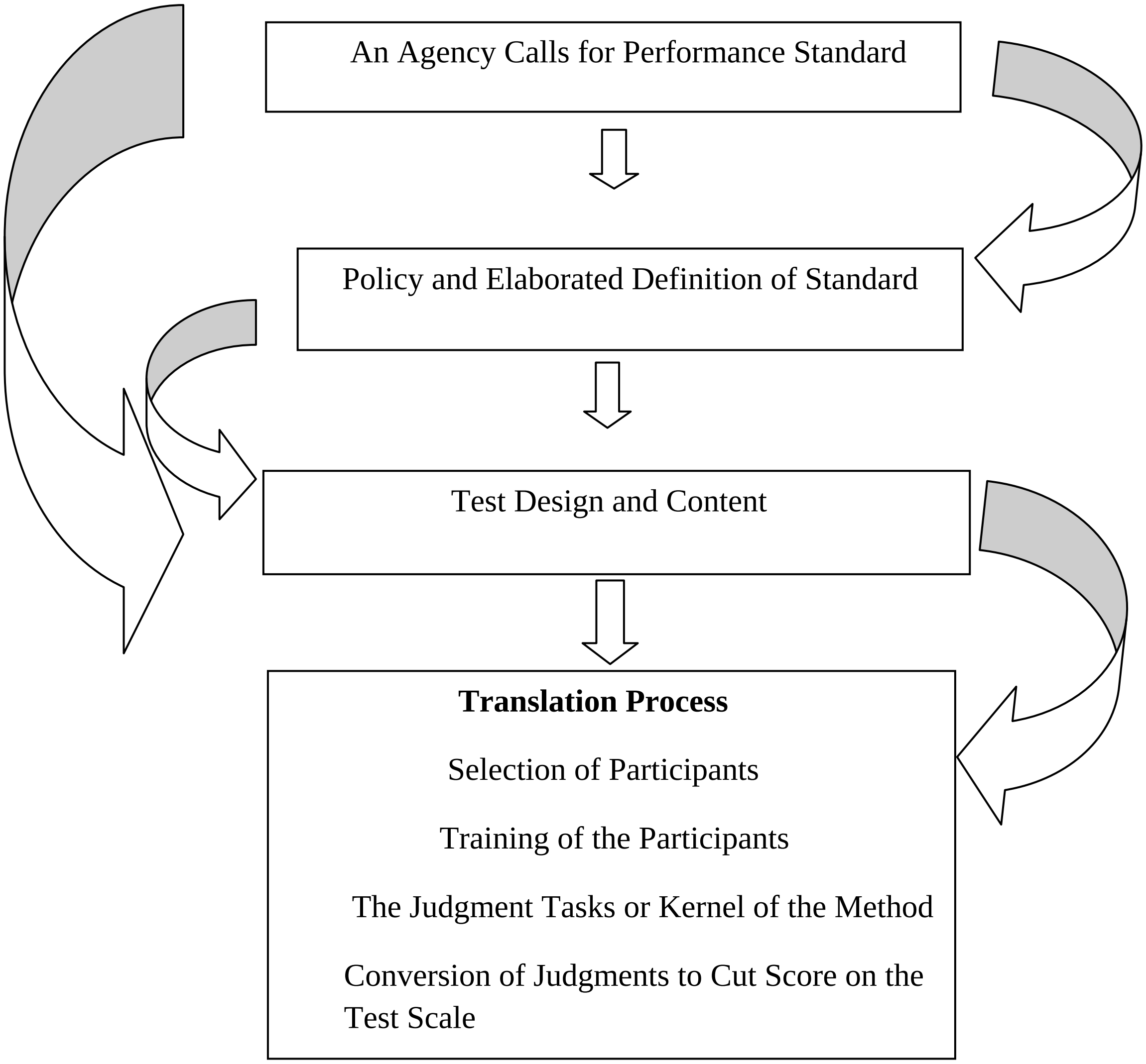
Figure 1-1: Relation Between Performance Standard and Cut Score (Proposed by Kane, 1994)


A. Performance continuum
x
Conceptual location on continuum of "minimally qualified" examinee (i.e. ability underlying the performance standard)
B. Test score continuum
y
Location of cut score

The standard translation process although the most essential part of standard setting, however is only the last operational part of the standard setting method. To reflect the full complexity of the standard setting process it is also necessary to establish the logical links between the entire operational components of standard setting. Specifically, the logical link between a performance standard, a test design, and standard translation methods adapted from Reckase (2009) are summarized in Figure 1-2. The arrows pointing downwards in Figure 1-2 shows that each subsequent step depends on previous ones, so that its result can be checked for consistency with the previous step.

Figure 1-2 begins with the policy setting processes that involve value judgments and which yields the performance standards. Both test design and standard translation constitute measurement operations. The performance standard serves as input in test design while both the performance standard and the test serve as inputs to the standard translation process. Standard translation encompasses both translation enabling operations of standard setting facilitators and actual translation process embarked on by the standard setting participants. It is deemed appropriate to provide the meaning of each of the standard translation operations. Hence, brief description of the standard translation operations is provided in Section 1.1.4., which immediately follows Figure 1-2.

Figure 1-2: The Parameter Estimation Model for Standard Setting (Proposed by Reckase, 2009)


An Agency Calls for Performance Standard
Policy and Elaborated Definition of Standard
Test Design and Content
Translation Process
Selection of Participants
Training of the Participants
The Judgment Tasks or Kernel of the Method
Conversion of Judgments to Cut Score on the Test Scale

1.1.4. Standard Translation Process

According to Reckase (2001), the typical operational components of standard translation in functional order include:

1) The selection of participants

2) The training of participants for the task of translation

3) A specific set of judgment tasks that is the kernel of the standard setting method

4) The provision of feedback and other supporting information to the participants

5) The conversion of the judgments to a reporting scale, and finally,

6) The reporting of the results of the process

These operational components of standard translation are discussed in the functional order in the following paragraphs of this section.

Standard translation requires participants that are fluent in both policy and test language (Reckase, 2001). Therefore in theory, participants should be selected for standard translation that possess the knowledge, skills, abilities, and other personality attributes desirable for performing the tasks (Raymond & Reid, 2001). However, in practice, it may not be feasible to find participants that possess all the cognitive competencies and personality attributes required for performing the tasks. As a consequence, measurement researchers and practitioners prescribe a specific set of judgment tasks, which Reckase (2001) called the kernel of the standard setting methods. They then strive to identify individuals who already possess many of the requisite knowledge and skills to perform the task and supplement for deficiencies in their knowledge and skill through training (Raymond & Reid, 2001).

The kernel of the standard setting methods are tasks prescribed to assist with the standard translation, most often judgmental tasks. According to Reckase, there is the tendency of standard setting researchers to label these tasks using short phrase. For example, they often indicate that a standard was set using the modified Angoff method or bookmark method "as if these methods were well-defined recipes for conducting a standard-setting process" (Reckase, 2001, p.160). Reckase highlighted that this practice of using short hand labels is extremely misleading because the phrase summarizes only the kernel of the method and ignores the wide variations in all of the other components of the standard setting process. In order to be consistent with the literature, the "kernel" would also be referred to for short as the standard setting method in this dissertation.

Apart from the kernel of the standard setting methods, arguably, most of other standard translation procedures are really training operations. However, little is known about training in the standard setting literature, a gap to be addressed by this dissertation. Due to sparse conceptualization of training in the standard setting literature, two of the three adopted definitions of training discussed in this paragraph were taken from the cognitive psychology literature. The definition of training taken from the standard setting literature was of an operation that is intended to modify participants behavior and that includes at least instruction, practice, and feedback while the definitions taken from the cognitive psychology literature were as follows: (1) any medium for exchange of information that could impact behavior and, (2) a planned effort to facilitate the acquisition of knowledge, skills, ability, and attitudinal behavior patterns required in order to perform adequately a given task (Latham, 1988; Wexley, 1984). These three definitions taken together implicated that training is a multi-faceted concept and a planned activity that involves exchange of information that is intended to influence cognitive behavior. Meanwhile, feedback is often represented as distinct from training. However, it was defined as information provided to guide the participants after at least one round of practice performing the translation task (Reckase, 2001). Because feedback is also intended to modify participant behavior it implies that feedback is part of the training (Raymond & Reid, 2001). As a consequence in Figure 1-2, feedback was not specified as a separate operation because it is assumed to be encompassed by training.

The last step of the standard translation process identified in Figure 1-2, involves the conversion of participant's judgments into cut scores. The task of converting participant's judgments to cut scores is the responsibility of the measurement experts. In a conventional standard translation process, the judgments of each participant are used to determine their individual cut score estimate. These cut score estimates are usually provided to them as part of feedback. The feedback usually is in the form of graphical displays of the cut score estimates and summary statistics for the group with discussion held around them. These feedback and information obtained from the discussion are meant to be used by the participants in the ensuing rounds of judgments and are typically directed towards fostering consensus among the participants. Because cut scores are part of the feedback provided to the participants, it was deemed necessary to describe also the cut score computation process as follows.

Cut scores for the test-centered methods notably, the Angoff method, are usually aggregated for all participants using the mean or median and are computed either on the IRT ability or true score scale (Wyse, 2009). However, for the purpose of this dissertation on the Angoff method, mean estimation methods are of interest and thus are discussed further. The judgments in the case of the Angoff method might comprise of probability ratings for a set of items. Cut score estimation for the Angoff method on the true score scales entail summing the probability estimates for each participant and averaging across participants for the group estimate. Cut score estimation on the IRT ability scales, assuming a unidimensional and a fixed underlying ability parameter for the Angoff method involve, mapping each probability judgment to an ability score, averaging the ability estimates over items and participants for the group estimate. IRT ability scales cut score estimation approach have been established to yield more accurate and unbiased cut score than estimation on the true score scale (Kane, 1987). However, for the purpose of this dissertation based on simplifying assumptions about the standard setting process, the computations are conducted on the true score scale. This concludes description of the standard translation process.

1.2. Problem Formulation

The importance of standard setting has been highlighted in recent years in the United States public school context of this dissertation. Since the year 2001, with the *NCLB public school accountability law* performance goal of 100 percent student proficiency in mathematics and reading by 2014, there has been increased use of standard setting procedures for deriving cut scores. These cut scores are used for determining adequate yearly progress (AYP) of public schools towards the goal of the NCLB law which statistics, are used for high stakes decisions, such as school staff dismissal and even public school closure (Chiang, 2009; Dillon, 2011; Figlio & Rouse, 2006; Jacob, 2005; Scott, Duffrin, Kelleher, & Neuman-Sheldon, 2009; Torre & Gwynne, 2009a, 2009b).

Advancement of the goal of the NCLB and other public school accountability initiatives is critically dependent on the efficacy of the cut scores in representing performance levels conceived by policy makers (Wyse, 2009). Some consequences associated with use of inaccurate cut scores and consequently AYP statistics in decision making with the NCLB law, for instance, includes: failing to identify schools needing intervention or, conversely, inappropriate identification of schools as requiring intervention, with the latter error potentially leading to external control, loss of employment by school staff, and school closure.

To make this discussion of consequences of use of erroneous cut scores with the NCLB law realistic, last year, 38,000 of the nation's 100,000 public schools fell short of the AYP under the federal rating system. Consequently, Arne Duncan proposed overriding the *NCLB* proficiency sanctions through waivers for schools implementing the *Race to the Top* alternative school improvement accountability initiatives (Dillon, 2011).

Despite the consequences associated with use of inaccurate cut scores for decision making, standard setting remains among the least understood areas of testing and psychometrics. For instance, cut scores continue to vary in mysterious ways across replications of standard setting studies (Cizek, 2001; Cizek & Bunch, 2007; Glass, 1978; Hambleton & Pitoniak, 2006; McGinty, 2005; Reckase, 2009; Shepard, Glaser, Linn, & Bohrnstedt, 1993; Skaggs & Hein, 2011; Wyse, 2009; Zieky, 2001). It may be uncontroversial to suggest that this state of affairs is due to sparse substantive theory guiding standard setting method and training research (See Reckase, 2001 also for this highlighted gap with feedback). This state of affairs persists even for the Angoff method, one of the first prescribed standard setting methods, and which has since emerged as been among the most researched and applied methods.

The Angoff method requires participants to judge conditional probabilities of correct responses to multiple choice test items for the target group of students who barely makes it in a performance category[1]. The Angoff method involves two cognitive tasks which in serial order are: (1) conceptualizing the student population that barely makes it in a performance category and, (2) estimating the proportion of the students who would answer each test item correctly (Hein & Skaggs, 2009, 2010; Impara & Plake, 1997, 1998; Raymond & Reid, 2001; Reid, 1991).

The Angoff method with task requiring conditional probability judgment, which in measurement terminology entails estimating item difficulties for the target student population on the probability metric has solid statistical and psychometric theoretical underpinning established to yield veridical outcomes when appropriately implemented. However, the Angoff method tasks were critiqued for been too cognitively complex for participants to execute in the context of

[1] The target groups of students are also referred to as borderline, minimally acceptable, minimally qualified, or minimally competent in the standard setting literature.

researching it with the National Educational Assessment Program (NEAP) public school standard setting as to undermine its usefulness for practical purposes (Shepard et al., 1993).

Since this critique of cognitive complexity of the Angoff method tasks was launched, it has stimulated a lot of research. However, the efforts to address this cognitive complexity by training have either introduced instructions for the first component task of conceptualizing the target group of students or feedback procedural modification and without substantive consideration of the knowledge and skills requirements of the tasks (Hein & Skaggs, 2010; Impara & Plake, 1997). Other efforts to address this cognitive complexity have introduced alternative standard setting methods that do not require participants to generate conditional probabilities and also without substantive consideration of the tradeoff between cognitive complexity of a task and the accuracy of outcomes (Hein & Skaggs, 2010; Impara & Plake, 1998).

Evidence provided by research addressing the cognitive complexity of the Angoff method tasks through introducing feedback procedures suggests that without feedback that participant’s judgment is flawed and that the current feedback approaches does help to remove inconsistencies. Therefore, the predominant conclusion has been that the prevailing feedback types are effective in addressing cognitive complexity of the Angoff method tasks (Brandon, 2004; Clauser et al., 2009a, Clauser, Mee, Baldwin, Margolis, &Dillon, 2009b; Clauser, Mee, & Margolis, 2011; Margolis, 2011; Mee, Clauser, & Margolis, 2011; Wyse, 2009; Wyse, in press). As a consequence, in the licensure and certification contexts where the Angoff method is still in use, implementation of the method has emphasized feedback (Cizek & Bunch, 2007). However, the prevailing feedbacks which are often provided without instruction on their meaning and on how to integrate them into judgment by no means reduce cognitive complexity of the tasks for the participants (Reckase, 2001). Moreover, they lead to less efficient Angoff standard setting in practical settings.

One of the alternative standard setting methods introduced consequent to the critique of cognitive complexity of the Angoff method is the Bookmark method (Mitzel, Lewis, Patz, & Green, 2001). The Bookmark method requires the participants instead to make only one decision namely, placing a bookmark on the first item in a test booklet of items ordered in terms of their item response model difficulty, that the students that barely makes it in a performance category cannot answer correctly with a specified probability (Hein & Skaggs, 2010). There is ample evidence to suggest that the Angoff method can yield more accurate cut scores than the Bookmark method (Cizek & Bunch, 2007; Haertel & Loriè, 2004; Reckase, 2006; Wyse, 2009). However, feasibility concerns seem to have swamped accuracy. The Bookmark method has now become the method of choice in the K-12 public school standard setting contexts for standard setting (Karontonis & Sireci, 2006).

Standard setting researchers recognize training as a multifaceted concept that includes instruction, practice, and feedback. However, little attempt has been made to understand the cognitive requirements of the Angoff method tasks and to apply them to the design of training (Raymond & Reid, 2001). One plausible explanation for the latter state of Angoff method training research affairs is the assumption that the participants understand the tasks and are able to perform them well (Wyse, 2009). As a consequence, the training interventions of instruction and practice have remained relatively unexplored as mechanisms for addressing cognitive complexity of the Angoff method tasks.

McGinty (2005) in recognition of the lacking cognitive theory supporting standard setting methods and training research called for the need for standard setting researchers to understand the “Black Box” of standard setting methods. The “Black Box” referred to the factors that drive the judgments of participants and the strategies underlying their judgments. Recent studies have started to heed McGinty’s call to understand the cognitive processes of the Angoff method tasks through introspective reports of research participants and have also focused on the first component task of conceptualizing the target group of students (e.g., Buckendahl, 2005; Ferdous & Plake, 2005; Giraud & Impara, 2005; Giraud, Impara, & Plake, 2005; Hein & Skaggs, 2010; Skaggs & Hein, 2011). However, the approach adopted by these studies to cognitive task analysis (CTA) is not the most productive way.

Consequently, this dissertation approach to addressing cognitive complexity of the Angoff method tasks is through CTA that drew from diverse knowledge bases that study the concepts and processes involved in the Angoff method tasks to understand its knowledge and skills requirements. Building on Raymond and Reid’s (2001) work, specifically, the CTA relied on measurement, statistics, and psychology knowledge bases, the crux of which was the probability judgment heuristic theories. The findings of CTA were applied to the design and evaluation of training. The new training addresses cognitive complexity of the Angoff method tasks by training with instructions, practice, and feedback interventions tailored to judgment heuristics principles that break up the task of judgment of conditional probabilities to simpler mental operations. Henceforth, the new training is called the Heuristic training while the abstracted prevailing training is called the Normative training.

The overriding goal of prescribing the Heuristic training program is to foster veridicality of laboratory standard setting research outcomes. This goal is to be accomplished by the Heuristic training through fostering conceptual and judgment strategy understanding. The secondary goal of the Heuristic training is to ensure efficiency of standard setting procedures.

The immediate goal of this dissertation empirical research is to test the effectiveness of the Heuristic training over the Normative training and the effectiveness of instructions and practice activities versus feedback procedures.

1.3. Empirical Research Purpose

By relaxing knowledge and skills assumptions in the design of the Heuristic training, much like a detective investigation, this dissertation empirical research is to find out where the source of challenge of participants lies in the execution of the Angoff method tasks. Specifically, the purpose of this dissertation's empirical research is to:

(1) Explore the Heuristic training outcomes for reasonableness of the modified Angoff heuristic training model assumptions

(2) Compare performance of participants of the Heuristic training to those of the Normative training in relation to the model assumptions and also based on internal and external validity criterion measures

1.4. Empirical Research Questions

The heuristic model and training methods are to be evaluated empirically. The evaluation approach is based on substantive considerations of the heuristic principle and content domain knowledge. The following three questions are of interest:

(1) How well does the Heuristic training recover heuristic model assumptions and underlying data structure compared to the Normative training?

(2) Does the Heuristic training perform better in increasing stability and correspondence of outcomes to reasonable substantive values compared to the Normative training?

(3) How and why does feedback work with the Heuristic and the Normative training, do they work due to substantive construct-relevant or construct-irrelevant factors?

1.5. Expectations

The ultimate expectation is that the qualitatively better Heuristic training instruction, practice, and feedback activities would yield best results in terms of increasing substantive meaningfulness of judgment outcomes, enhancing judgment accuracy, and efficiency than the Normative training approaches. The immediate testable expectations are that the Heuristic training based on adequate attention to the knowledge and skills requirements of the Angoff tasks in the design of all facets of training of instruction, practice, and feedback would result in the following outcomes:

(1) Promote consideration of knowledge and skills constructs measured by the tests in judgments

(2) Increase consideration of and integration of experiential information and the knowledge and skills constructs measured by the tests in judgments.

(3) Enhance substantive meaningfulness of judgment process, technical qualities of reliability and validity of item difficulty judgments, and reasonableness of cut scores

1.6. Overview of Dissertation Chapters

The aim of this section is to give the reader a preview of what is to come in the rest of the chapters. This dissertation is comprised of eight chapters. Chapter one served to introduce standard setting, the topic of this dissertation, and the problem to be investigated by empirical study. Chapter two follows from Chapter one and presents the measurement, statistics, and psychology theoretical foundations of this dissertation research on the Angoff standard setting method. Chapter three reviews research on the Angoff method. Chapter four presents reformulations of the Angoff method based on the preferred theoretical positions and in terms of posited cognitive and non-cognitive factors influencing the probability judgment task. Chapter five considers the evaluation and analytic framework for the training and probability judgment outcomes. Chapter six delineates the methods of this dissertation’s empirical studies. Chapter seven elaborates on the results of empirical studies. In Chapter eight the background of the study, intellectual merits of dissertation, findings, interpretations of findings, recommendation for practitioners, limitations and directions for future research are discussed.

Chapter Two: Theoretical Framework

This chapter reviews theories in the research fields of measurement, statistics, and psychology that grounded the CTA and training design for the Angoff method. There are four major sections in this chapter, each reviewing the relevant theories that contributed to understanding of the cognitive task requirements of the Angoff standard setting method. The first section reviews measurement theories; the second section reviews probability theories; the third section reviews cognitive psychology theories; and, the fourth section reviews social psychology theories. It turns out that there have been many ideas expressed in the research fields of measurement, statistics, and psychology that are directly relevant to the study of the Angoff standard setting method. Rather than just present my preferred positions for this dissertation in isolation, as you will see in Chapter four, it is appropriate to review also the various dominant positions on each of the topics and their relationships. However, if the reader is impatient with such discussion, it is possible to skip to Chapter three which reviews the Angoff method literature to which the preferred theoretical views reviewed directly generalizes.

2.1. Theoretical Framework for Measurement

As mentioned in the introduction, standard setting is an educational as well as a psychological measurement problem. Therefore, it is appropriate to provide a working definition for psychological measurement. Psychological measurement was defined as the systematic way of assigning numbers to individuals as a means to represent their studied characteristics (Raykov & Marcoulides, 2011). By and large, psychological measurement is concerned with individual attributes however most of the attributes of interest are unlike physical attributes (e.g. height) which can be measured directly. The immediately following sub-section briefly describes the attributes of interest in psychological measurement.

2.1.1. What to Measure

Individual attributes of interest in psychological measurement are referred to as constructs. They are referred thus because they cannot be measured directly. These attributes are abstracted from observable behaviors. The utility of these constructs in behavioral sciences is that they help us to classify individual atomistic behavior thereby facilitating substantial reduction of behaviors (Raykov & Marcoulides, 2011). Therefore the atomistic behaviors are fallible indicators of the constructs. For instance, performance standards measured by standard setting procedures are expressed in terms of knowledge, skills, and abilities constructs, while test items are presumed to be fallible indicators of these knowledge, skills, and abilities constructs. Because performance standards are measured by standard setting procedures, it is appropriate to review measurement theories. The immediately following sub-section presents theoretical views about measurement.

2.1.2. How to Measure

The purpose of this section is to briefly review and compare and contrast conceptual views about how to measure. There are two views on how measurement can be accomplished, referred to as the realist and the operational views (Mari, 2005). The realist view considers science as the study of the external world independent of the observer whose experiments and observations are simply means of finding out about the world. Consequently, realists construe of quantities as properties which exist prior to measurement thereby explicating a theoretical and observational dichotomy. The theoretical terms play causal roles and serve to organize, explain, and predict data (Dingle, 1950). For instance, in the context of standard setting, the performance standards represent the theoretical realm so that they serve to explain the cut scores estimates which represent the observational part of the dichotomy. The realist view is classified further as classical or representational. What these two sub-categories of the realist view share is the metaphysical assumption that there is a true value of a quantity which plays a fundamental role for the measurement science. However, they are distinguished in that while the classical view assumes an empirical theory of numbers, the representational view makes no commitment to numbers as empirical entities and to the assumption that measurement is a process for discovering quantitative attributes (e.g. order and additivity). Put differently, the representational theory opposes the idea that numbers are in the world rather it maintains that we assign numbers to nature (Hand, 1996; Mari, 2005; Michell, 1990, 1999).

According to the operational view, measurement is any precisely defined set of operations that yields a number. In other words, the operationalists construe measurement operations as giving sense to quantities. Steven's (1946) definition of measurement as a process of assigning numbers to objects in accordance with rules is consistent with this operational perspective. Steven's definition of measurement was controversial and was critiqued for the following reasons: (1) obscuring the distinction between representational rules; (2) failing to highlight that it is the attributes of the objects and not the objects themselves that is being measured; and, (3) for widening the concept of measurement that only excludes random assignment from its purview.

Despite the shortcomings and controversies about Steven's definition of measurement, it remains the preferred notion of measurement among psychologists and by extension to standard setting. Consequently, it ushered a floodgate of measurements because, almost anyone could devise rules to assign some numbers; it shifted the focus of measurement from numerical facts to rules for making numerical assignments; and, from quantitative attributes to objects and events (Suen, 1990). The realist versus operational views of measurement also drive standard setting research debates with the operational being the contemporary view of standard setting researchers. However, the realist view of measurement is the preferred notion of measurement in this dissertation. However, regardless of the preferred view of measurement, there are different qualities of numbers that can emerge from the measurement process. Therefore, the immediately following section reviews the different types of measurement numbers.

2.1.3. Types of Measures

The operational definition of measurement was not the lone contribution of Steven's to measurement theory. Steven (1946) also described four ordered levels at which measurement can occur namely: nominal, ordinal, interval, and ratio. These levels differ in terms of quality of numeric information they represent. The desirable qualities of numeric information in order of ascendancy are order, magnitude, equality of interval, and absolute zero.

The nominal level is the lowest level of measurement at which the assigned number simply denotes a naming or identification scheme; the ordinal level is the second level of measurement at which order and magnitude are represented in the set of data; the interval level is the third level of measurement at which the properties represented are equal intervals and absolute zero of differences between the points; the ratio level is the fourth level of measurement at which the measurement numbers possess all four properties (i.e. order, magnitude, equality of interval and absolute zero).

According to Steven, the quality of the information represented by measurement numbers should be taken into consideration in deciding on appropriate statistical method for data analysis. The question then is how do we know the quality of information represented in measurement numbers? The approaches to determining the quality of information represented in measured numbers are briefly reviewed in the immediately following section.

2.1.4. How to Evaluate Measures

There are two fundamental concepts underlying the quality of measurement numbers and of measures of constructs: reliability and validity. Reliability addresses consistency of measures and evaluates how much of the true quantity of interest, for example threshold proficient ability, is contained in the observed measure. Validity is a multifaceted concept. There are three main types of validity namely: content, criterion, and construct validity. Construct validity is the more encompassing of all three and, refers to the extent to which explanatory concepts account for performance on test and to which test scores recover hypothesized relationships amongst constructs in the substantive area (Raykov & Marcoulides, 2011). Hence forth all mention of validity in this dissertation refers to construct validity.

It is also important to highlight some relationships, similarities, and differences between the evaluation concepts of reliability and validity. Reliability and validity are inter-dependent so that they can be viewed as two regions on the same continuum (Raykov & Marcoulides, 2011). On the other hand, while evaluation of reliability of measures focuses more on their statistical properties, evaluation of validity of measures focuses more on their substantive properties. For example, with the reliability criterion, estimated item difficulties and cut scores are evaluated positively if they contain no more than negligible error with error measured by applying generalized linear modeling statistical framework. On the other hand, evaluation of estimated item difficulties and cut scores based on the validity criteria would require investigating their meaningfulness with regards to how well they recover a priori hypothesized relationships between the knowledge, skills, and abilities constructs. Also, making validity argument requires collecting evidence to support the inferences (e.g. pass/fail decisions) that are to be drawn based on the outcomes of measurement process (Messick, 1989).

2.2. Theoretical Framework for Probability

This section provides an overview of theories of probability that informed this dissertation research. The concept probability is used in various contexts including situations further away from the typical chance events (Sun, 2003). The question becomes, are these uses of the same meaning? This question pertains to the problem of interpretation of the concept of probability. In his paper, Popper (1959) gives the meaning of interpretation of probability as the interpretation of statements, $p(y, x) = r$ which reads, the probability of y given x is equal to r; where r is a real number, y represents the event; and x is the conditions of the experiment.

2.2.1. Interpretations of Probability

Two major interpretations of probability stand out namely: the objective and the subjective distinctions (Hays, 1994; Howell, 2002; Kerlinger, 1986; Popper, 1959; Sun, 2003). The subjective interpretation simply regards probability as a way of dealing with partial knowledge so that $p\ (y, x) = r$ is a measure of degree of rational belief which the information x gives about y. On the other hand, the objective interpretations are characterized by interpreting $p(y, x) = r$ as a statement that can, be objectively tested, using statistical tests consisting of sequences of experiments. In the objective realm, x in $p(y, x) =r$ refers to the experimental conditions; while y describes some of the possible outcomes of the experiments; and the number r describes the relative frequency with which the outcomes y is estimated to occur in any sufficiently long sequence of experiments characterized by the experimental conditions x (Popper, 1959).

Even though the objective and the subjective interpretations are treated more or less the same way mathematically, however, critics are of the view that subjective probability is unreliable (Hays, 1994). In particular, subjective probability may vary from individual to individual for the same event because individuals vary in their backgrounds, knowledge, available information, and degree of belief about the occurrence of an event. Also, while the objective probability interpretation is emphasized in statistics, the subjective interpretation is emphasized in the cognitive psychology research area of probability judgment (Gigerenzer, 1994).

According to Beach and Braun (1994), the year 1967 marked the emergence of psychology as the official home for research on subjective probability. In that year, three publications came out addressing different aspects of most of what was known at that time both theoretically and empirically about subjective probability. Since then, series of research paradigms based on the notion of subjective probability have emerged a couple of the most prominent of which would be reviewed in the next section on the cognitive psychology theories. However, preference for the realist notion of measurement that views measurement as a process capable of recovering a population parameter necessarily follows preference of the objective view of probability. Therefore, reviewed further in the immediately following sub-section are theories of the objective probability schools of thought.

2.2.2. Objective Probability

There are two major types of objective probability interpretations namely: the propensity and the relative frequency interpretation (Popper, 1959). According to the frequentist interpretation, probability is the property of a sequence, precisely, the relative frequency of a long series. Therefore, it is possible to speak about probabilities only in reference to a properly defined sequence which was called a collective. A collective is defined as an unlimited sequence of observations which defines probability under two conditions: if relative frequencies of particular attributes within the collective tend to fixed limits and the fixed limits are not affected by any place selection. These were called the principle of convergence and the principle of randomness, respectively (Sun, 2003). However, the relative frequency interpretation is limited in the following two ways: (1) because of its reference to infinite sequences of limiting values, it is unsuitable for interpretation of single case probabilities; and, (2) there are infinite possible choice of reference class for an event, and it is ambiguous to determine the homogeneity of a sequence. Reichenbach (1949) attempted to address the problem of choice of reference class by suggesting that the appropriate class is the narrowest class for which reliable statistics can be compiled. However, this modification was inadequate since some individual cases do not seem to fall into any existing classes and besides, the limit of relative frequencies in narrowing classes may not exist.

Due to the aforementioned limitations of the frequentist probability interpretation, the propensity interpretation was proposed, which refers instead to relational properties of experimental arrangements or conditions kept constant for experimental repetition. The propensity interpretation set forth the hypothesis that every experimental arrangement generates physical propensities which can be tested by frequencies. The concept propensity is necessarily a hidden property which is not directly observable but it is this property that brings about the result of relative frequencies which can be viewed as its outward expressions. The propensity interpretation is an improvement over the frequentist as it gave insight that probability may be more closely related with the generalizing conditions than with the actual collective generated from an experimental arrangement. Hence, it was a useful conception of probability because it drew attention to unobservable dispositional properties of the physical world and was relevant for interpretation of physical theory. Consequently, Popper (1959) and Sun (2003) argue for the propensity interpretation primarily because of the problem of interpreting probability of singular events (occurrences). In essence, by conceiving of propensity as a property of the generating conditions, it facilitated interpreting probability of singular events. Since propensity treats probability as an objective and physical property, it is useful for predicting future events. However, the challenge is that unlike ordinary physical concepts, the measurement of propensity requires an indefinite number of experiments. Nevertheless, this shortcoming is not enough to deter from its theoretical importance.

The preferred notion of probability in this dissertation is as being on an objective or subjective continuum rather than a dichotomy. This view implicates that there are different degrees of subjectivity reflected in measured probabilities. However, the aspiration is for elicitation of purely objective probability estimates from participants. To achieve the latter state of affairs, the speculation in this dissertation is that the original goal of prescribing the Angoff method was not to study states of the minds or changes in degrees of belief of individuals as new information is gained (Hays, 1994; Sun, 2003). It is also assumed that classroom teachers in the public school contexts conduct experiments if you will, of repeated testing of students to measure their abilities that yield empirical facts. Consequently, the objective probability interpretation is preferred as the more suitable for the Angoff standard setting method because, it is apt for experimental contexts where the objective conditions of the events are well-defined and reproducible.

2.3. Theoretical Framework for Cognitive Psychology

This section presents the cognitive psychology theories that were relevant to this dissertation. Cognitive psychology theories were the crux of the re-conceptualization of the Angoff method in this dissertation, consequently this section is more elongated than other sections of this chapter. Specifically, there are four subsections in this section. The first sub-section reviews the concept of judgment; the second sub-section reviews judgment process theories; the third sub-section reviews probability judgment heuristic research paradigms; and, the fourth sub-section reviews fundamental cognitive processes underlying the representative and the availability heuristics that were applied to the reformulation of the Angoff method.

2.3.1. Judgment

Cognitive processes are often described in terms of degree of consciousness about rules applied in performing a task. Consciousness refers to a person’s ability to report from the working memory about how a task was accomplished (Gigerenzer & Murray 1987). The precise cognitive process of interest in this dissertation is judgment. Therefore, a brief description of the notion of judgment is appropriate. The psychology of judgment was conceived of as sharing features with the psychology of perception and the psychology of thinking because, judgment can be slow and deliberate like problem solving, or quick and immediate like perception (Keren & Teigen, 2004). The latter conception of judgment highlights that it can sometimes be an unconscious or conscious process. However, the preferred notion of judgment for the purpose of this dissertation is as a slow and deliberate thinking process of integrating information into probability judgments (Keren & Teigen, 2004; Slovic & Lichtenstein, 1971).

It is important to dissociate judgment from decision making because they are often used interchangeably. The following differences were highlighted by Eysenck and Keane (2010) between judgment and decision making: (1) while judgment research focuses on understanding how people use different cues to make inference about situations and events, decision researchers address how people choose different courses of action to achieve their goals (Hastie, 2001); (2) judgments are evaluated in terms of their accuracy while decisions are evaluated in terms of the consequences of the chosen actions (Harvey, 2001); (3) as a type of problem solving judgment requires generating options while in contrast in decision making the options are generally present to the problem solver; and, (4) decision making is typically concerned with preferences while judgment is typically concerned with solutions. In addition, decision making was conceived of as the broader framework of thinking with judgment pertaining to those aspects of the process that are concerned with estimating the likelihood of various events (Eysenck & Keane, 2010). Therefore, research on judgment is focused on understanding the cognitive processes underlying probability judgments, also known as judgment of uncertainty. Having perhaps clarified the meaning of the concept of judgment as applied in this dissertation, the next logical step is to review the theories of judgment in accordance with this conception. Therefore, the immediately following section reviews theories of how judgment is made.

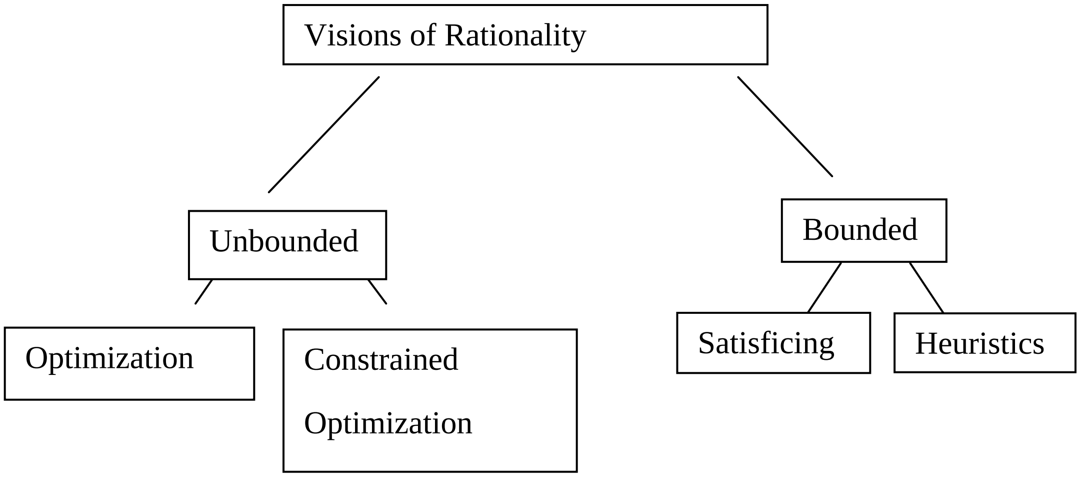
2.3.2. Judgment Process: Theories

There are three theories about how human beings make judgments namely: the normative, the prescriptive, and the descriptive theories. The normative theories posit how human beings should reason in making judgments. Examples of normative theories are formal logic, probability, and decision theories. The normative models of thinking embody idealized, rational, and unlimited memory capacity principles that are descriptive of how the super-intelligent people make judgments. They emanate through the processes of reflection and analysis and are the duty of philosophy (Baron, 2004; Over, 2004). The normative models of judgments often serve as the standards for evaluating judgments and include models developed in the probability and statistics areas of study. On the other hand the descriptive theories elaborate on how the average persons, also referred as ordinary people actually reason and take into account both actual behavior and reflective judgments.

The prescriptive theories ideally are the product of the normative and the descriptive theories and recommendations of the type of reasoning to employ in different judgment contexts to approximate the normative ideal (Baron, 2000). Therefore the goal of prescription should be to improve human judgment which can be accomplished through devising ways of reasoning that could reduce discrepancies of judgments from the normative ideal. These prescriptions for correcting biases in human judgments, with bias defined in relation to the normative ideals are called prescriptive models (Baron, 2004; Over, 2004). The task of prescription rests on applied fields of study that include education likewise, this dissertation is devoted to the task of prescription and drawing from both the normative and the descriptive judgment theories.

Meanwhile, because of the conception of judgment as a deliberate thinking process it is appropriate to provide a little bit of background about the psychology of reasoning. The study of probability judgment is often concerned with understanding the reasoning processes involved in performing the task which results are often used to infer rationality of humans. There are two broad views about human rationality: the bounded and unbounded views of rationality (Gigerenzer, Todd, & ABC group, 1999). These two views and their sub-types are summarized in Figure 2-1 that follows:

Figure 2-1: Visions of Rationality (Adapted From Gigerenzer et al., 1999)



The notion of unbounded and bounded rationality originated from Simon (1957). The unbounded view of rationality assumes that all relevant information required for judgment is available and that the human mind has limitless reasoning capacity. Unbounded rationality models include optimization probability models (e.g. Bayesian). These models do not take into account the constraints of time, knowledge, and computational ability that humans are faced with in the real world (Gigerenzer et al., 1999). The unbounded view of human rationality assumes that human judgment is the product of optimization process so that the best judgment is made.

However, Simon believed instead that humans possess bounded rationality. The notion of bounded rationality means that human reasoning is constrained by the environment or by limited processing ability of the mind. Constraints from the environment include availability of information or time while constraints in the mind are either due to limited knowledge, attention, memory, and computational resources. Examples of bounded rationality view include satisficing and heuristics. Satisticing is a cognitive strategy restricted to the decision making task of searching through a sequence of available alternatives while heuristics have wider generality. The focus of this dissertation is on the heuristic notion of bounded rationality. Therefore, the rest of the discussion in this section pertains to heuristics.

To summarize there were three parts to the principle of bounded rationality laid out by Simon (1957) namely: (1) limits in processing capacity warrants the use of approximate methods like heuristics in performing most tasks; (2) application of different approximations in performing tasks yields different solutions; and, (3) to describe, predict, and explain behavior it is essential to construct theories of heuristic approximations and to describe the environments for which they are suitable. According to the Simon's bounded account of rationality, although human beings are faced with constraints in reasoning they are still able to produce reasonable solutions to problems by using various short-cut strategies which are referred to as heuristics. Perhaps, it is appropriate at this point to provide also a little bit of historical account of heuristics.

According to Gigerenzer et al. (1999), the term heuristic is of Greek origin and means serving to find out or discover. Since its adaptation to English many definitions of heuristics exist. For the purpose of this dissertation a modern definition of heuristic is preferred. The preferred notion of heuristic is "a strategy that ignores part of the information with the goal of making judgment more quickly, frugally, and accurately than more complex methods" (Gigerenzer & Gaissmaier, 2011, p. 454). This notion of heuristic is preferred for this dissertation because it makes more logical sense for laboratory Angoff standard setting method research paradigm. As you will see later in Chapter three, one criticism of the Angoff method is that it is too cognitively complex for the participants so that addressing this problem requires a heuristic prescription for the task that reduces cognitive effort through use of less information. Consequently, all review of the cognitive psychology probability judgment research paradigm pertains to heuristic cognitive processes that use little information in judgments. The section that immediately follows reviews two of the predominant research programs on probability judgment heuristics.

2.3.3. Probability Judgment: Heuristics Research Paradigms

There are two predominant research programs on probability judgment heuristics namely: the heuristics and biases and the fast and frugal heuristics programs. These two research programs are based on contrasting views about the utility of probability judgment heuristic strategies. To highlight some of the philosophical differences between these two research traditions, perhaps a chronological review of the research programs would help and begins in the immediately following paragraph.

In the 1970s, while artificial intelligence researchers were glorifying heuristics for their potential of making computers smart, in contrast, psychologists of the heuristics and biases program (Tversky & Kahneman, 1974), and of the two-system theories of reasoning (Evans, 2008; Gigerenzer & Gaissmaier, 2011) focused on demonstrating human reasoning errors, which errors they explained away as occurring due to the application of heuristics in judgment. In other words, in accordance with the heuristics and biases and the two-system research programs, heuristics were explanatory constructs of why human judgment was fallacious while content-free laws of logic and probability were glorified as embodying principles of sound thinking (Kahneman & Tversky 1972,1982; Tversky & Kahneman 1971,1973, 1974). In concrete terms, experimental results in the heuristics and biases tradition were typically interpreted as indicating some kind of human error usually attributed to one of three heuristics: representativeness, availability, or anchoring and adjustment (Tversky & Kahneman, 1974). An error associated with the representativeness heuristic for instance, is the conjunction fallacy which refers to violation of the probability axiom that the conjunction of two events is less than the probability of either event. This violation is spotted based on the subjective conception of probability by comparing judgment to the Bayes rule.

Evidence from the heuristics and biases program suggested that ordinary people use little information and apply limited cognition and are largely unable to estimate probabilities. The proponents of the heuristics and biases program even argued that since experts are prone to similar mistakes that it might be best to exclude the general public from making important judgments. Citing a Newsweek article reporting on heuristics and biases research, Gigerenzer et al. (1999) described most people “as woefully muddled information processors who stumble along ill-chosen short-cuts to reach bad conclusions” (McCormick, 1987, p. 24). Given this pessimistic view about capacity of humans to make probability judgment, it was hard to know where to turn to for reasonable judgments.

Meanwhile, in order to ensure cohesiveness of the ensuing discussions, the major contributions of the heuristics and biases program, namely the representativeness and the availability heuristics are briefly reviewed before formal description of the fast and frugal heuristic perspective. It is appropriate to review the representativeness and the availability heuristics because they were directly applied to understanding the cognitive processes and for formulating a model for the Angoff standard setting method task. The representativeness and the availability heuristics were identified as mediating most intuitive judgments of probability in many different contexts (Tversky & Kahneman, 1971).

The representativeness heuristic mediates probability judgment through decision processes of what is similar. When used by people, the representativeness heuristic reportedly leads to assigning events which are typical of a class a high probability of occurrence (Eysenck & Keane, 2010). The representativeness heuristic is typically used when people judge the probability that an object or event belongs to a class or process. For example, given a description of an individual, the probability that the individual is representative of an occupation would be determined by the similarity between the description and the stereotype of the profession. For the purpose of implementation, representativeness is incomplete without specifying the dimensions along which similarity is to be judged (Wallstein, 1983). The availability heuristic mediates judgment of probability of events through ease of retrieval of information from the long-term memory. The availability heuristic mediates judgment of probability through decision processes of what comes easily to mind (Tversky & Kahneman, 1974). In comparison, while the representative heuristic estimates probability by assessing similarity or connotative distance, availability estimates probability by assessing associative distance (Tversky & Kahneman, 1973).

Now to the fast and frugal perspective of heuristics, in contrast to the heuristic and biases approach to study of heuristics, it ushered in positive research accounts of heuristics. Subscribing to Simon's bounded rationality and to Brunswik (1955) ecological rationality view, Gigerenzer et al. (1999), a group of researchers on the adaptive behavior of thought (acronym ABC group), as opposed to seeing heuristics as often leading humans to make errors of judgment argue that heuristics are often very valuable tools for accurate judgment. Their central focus was on fast and frugal heuristics which involve rapid processing of relatively little information. The fast and frugal perspective assumed that humans possess an “adaptive toolbox” consisting of several such heuristics. Gigerenzer and his colleagues argued that fast and frugal heuristics, despite their simplicity, can be surprisingly effective and useful for making rapid judgments. Besides, according to this perspective, individuals with little knowledge when applying the fast and frugal heuristics can sometimes outperform those with greater knowledge.

According to this fast and frugal heuristic perspective, logic and probability are not good descriptions of how actual people make decision in the world including experts on judgment. A story told by Gigerenzer and Gaissmaier (2011), to underscore the latter point, was about an expert on judgment who was struggling with the decision of whether to leave or to stay in his current job, when called aside by a fellow professor and asked, why not "maximize your expected utility," you always write about that? The expert exasperated responded "Come on this is serious." Also, unlike the heuristics and biases research tradition, researchers studying heuristics from the fast and frugal perspective advocated for building theories and models for heuristics that reaches beyond a list of labels of heuristics. Research efforts directed at this model building shed light on the common building blocks of heuristics (Gigerenzer et al., 1999).

To summarize the philosophical differences between these two research traditions, Gigerenzer, et al. (1999) highlighted the following three fundamental differences: (1) the fast and frugal perspective views heuristics as a way the human mind can take advantage of the structure of information in the environment to arrive at reasonable decisions while the heuristics and biases approach views heuristics as unreliable aids that the limited human mind relies often on despite their inferior performance; (2) the heuristics and biases perspective uses the laws of probability as yardsticks of inferring rationality, this criteria is called coherence criteria, while the fast and frugal perspective uses more encompassing correspondence and ecological rationality criteria (e.g. of criteria; accuracy, speed, and frugality); and, (3) the fast and frugal perspective opts for computational models of heuristics while the heuristics and biases approach uses instead vague labels which because are unspecified can be fit to almost any empirical result post hoc.

Notably, although the fast and frugal perspective yielded less research, it offered great insights about the nature of heuristics; on the other hand three decades of prolific research by the heuristics and biases program generated only indistinct proposals of simple mechanisms of reasoning (Gigerenzer, 1996; Gigerenzer, 2008). To support this anecdotal claim was empirical research evidence that showed that the simple heuristics were more accurate than standard statistical models that had the same or more information when they were formalized. This result became known as less-is-more effects and put heuristics on par with standard statistical models of rational cognition (Gigerenzer, 2008). Perhaps the reader is wandering how all this review relates with Angoff standard setting. If that is the case, I plead that you bear with me as the immediately following paragraphs begins to make these connections.

Reading through the heuristics and biases and the fast and frugal perspectives approaches to studying probability judgment one could not help but notice the striking semblance between the heuristics and biases approach and the current approach to standard setting research. To highlight but three of these striking similarities: (1) just as heuristics and biases researchers use labels for heuristics, the standard setting researchers are also in the habit of using labels to summarize judgmental processes; (2) although standard setting research has advanced methodologically researchers still complain about the elusive nature of the enterprise (please see Chapter three for attacks on Angoff standard setting method) and, (3) the most striking semblance, the one that actually motivated this work is that standard setting research is not theory driven.

On this note, it was important to also highlight that although the heuristics and biases and the fast and frugal heuristics research programs employed different strategies to the study of heuristics, however they both highlighted the role that simple psychological heuristics play in human thought (see, Gigerenzer et al., 1999; Tversky & Kahneman, 1974). Also, apart from suggesting fundamental heuristic principles employed in judgment, both research approaches have strengths and weaknesses which standard setting researchers can certainly build on in studying Angoff standard setting method.

For example, some prominent ones identified by Eysenck and Keane (2010) for the heuristics and biases approach were as follows: (1) heuristics are used in many different ways by different researchers so that there is a danger of losing most of its meaning; (2) they fail to account for the fact that some errors of judgment occur because participants misunderstand the problem; (3) the emphasis has been on the notion that people’s judgments are biased and error-prone because information is processed in a biased way, which often seems unfair considering that humans also make incorrect judgments because of poor quality of available information; and, (4) much of the research is artificial and detached from the realities of everyday life, for instance although emotional and motivational factors play a role in the real world but they were rarely studied in the laboratory as a result, it is hard to generalize from laboratory findings (Lerner, Gonzalez, Small, & Fischoff, 2005).

On the other hand, some of the limitations of the fast and frugal heuristics were that: (1) they are used much less often than predicted theoretically, and some of the heuristics were by no means as simple as Gigerenzer and others have claimed, (2) they de-emphasize the importance of the judgment in question; and, (3) unless the conditions under which certain heuristics would be selected over others can be specified, their predictive and explanatory power remains questionable. Last but not least, there are ample results that standard setting research can draw from these research areas that can generalize to the Angoff task of judging probabilities.

For instance, Eysenck and Keane (2010) identified some research evidence that can inform research on the Angoff method. Some prominent ones that are particularly relevant for understanding and designing Angoff method training and feedback process include: estimates of probability change in the light of new evidence and Bayes rule formalizes some strategies to quantify these revisions; according to support theory, the subjective probability of an event increases as the description of an event becomes more explicit and detailed; according to Gigerenzer and Hoffrage (1995) judgments are more accurate when they are based on natural sampling and frequencies rather than probabilities albeit people often adopt biased sampling strategies and are inaccurate even when using frequency data, etc.

To conclude this section of review of research on probability judgment heuristics, the representativeness and the availability heuristics were identified as mediating most intuitive judgments of probability in many different contexts (Tversky & Kahneman, 1971, 1974). It is assumed that the representativeness and the availability heuristics would extend to the Angoff task. Going beyond the representativeness and the availability heuristics labels, categorization and memory were identified as the fundamental cognitive processes underlying the representativeness and the availability heuristics, respectively and were used to explicate the model for the Angoff task. As a consequence, it is essential to also review categorization and memory theories. The sub-section that immediately follow briefly sheds light on theories and principles underlying memory and categorization as generalized to the Angoff standard setting task.

2.3.4. Cognitive Processes of the Representativeness and the Availability Heuristics

This section provides a snap shot review of memory and categorization theories.

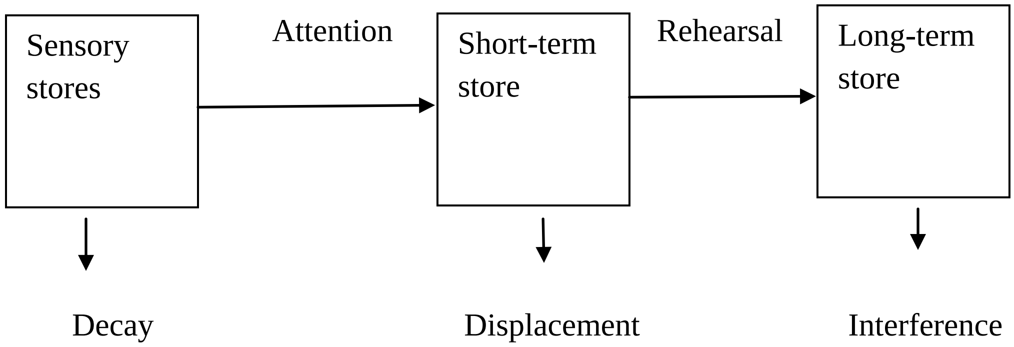
A. Memory: Concepts and Theories

In every day discourse, we often make reference to memory of past occurrence. This type of memory is what is called the retrospective memory in the memory research literature. There is another equally important type of memory called the prospective memory, which is memory for future commitments (Eysenck & Keane, 2010). Although these two types of memory are equally important in order to thrive in real life endeavors, for the purpose of laboratory Angoff standard setting research, however, retrospective memory is the more relevant. The proposition is that laboratory Angoff standard setting tasks are retrospective memory tasks because they require participants to reference their past experiences in making predictions about the future. As a consequence of the latter proposition, this review is restricted to retrospective memory.

Meanwhile, it seems appropriate to begin this review with a working definition of retrospective memory. Retrospective memory was defined as the processes involved in retaining, retrieving, and using information after the original information is no longer present. Information includes stimuli, images, events, ideas, and skills, etc. (Goldstein, 2008). Theories of memory generally consider both the architecture of the memory system and the processes operating within the system. While architecture refers to the structure of the memory system, processes pertain to the activities occurring within the memory system **(**Eysenck & Keane, 2010; Jenkins, 1979; Roediger, 2008). Even though the focus of this dissertation is on the processes, because these processes occur within memory systems, it was deemed necessary to review also the organizational scheme of the memory.

The traditional view of memory was a multi-storage system comprised of three storage systems namely: sensory, short-term, and long-term memory systems. This nascent view was at the heart of the early model of memory called the multi-store models (Atkinson & Shiffrin, 1968). The multi-store memory view was adopted for this dissertation so that all memory research reviewed in the following paragraphs was conducted in accordance with this perspective. The multi-store model was conceived of in accordance with the information processing theory (Broadbent, 1958; Waugh & Norman, 1965). Consequently, the model suggests that human memory is a lot like the computer's processing system so that for instance, remembering an event requires that one gets information into the brain, retains that information, so that later on when needed it can be retrieved and reported back (Atkinson & Shiffrin's, 1968; Myers, 2004). According to this multi-store model, the role of attention and rehearsal is to control flow of information between the memory stores while decay, displacement, and interference were the posited processes through which information are lost from the involved memory systems (Goldstein, 2008). The multi-store model is presented in Figure 2-2 that immediately follows.

Figure 2-2: The Modal Multi-Store Model of Memory
(Proposed by Atkinson & Shiffrin, 1968)



The multi-store model assumes that there is an important distinction between the capacity to briefly remember say a telephone number and the capacity to remember psychological theories (Eysenck & Keane, 2010). While the capacity to temporarily retain information is called short-term memory, the capacity to retain information over long period of time is called long-term memory. There is ample research evidence to support this distinction between short-term and long-term memory. For example, existing evidence indicate that these two systems differ in terms of temporal duration, storage capacity, forgetting mechanisms, and effects of brain damage (Eysenck & Keane, 2010; Keppel & Underwood, 1962; Miller, 1956; Murdoch, 1962; Petersen & Petersen, 1959; Sperling, 1960).

A classic experiment was conducted by Murdoch (1962) that distinguished between the notion of short-term memory and long-term memory. The experiment asked participants to read stimulus list material and at the end to write down how many words they could recall. The finding from the experiment based on plotting of the proportion of a large number of participants that recalled a stimulus item as a function of its serial position on the list showed a function called the serial-position curve. According to Goldstein (2008), Murdoch's serial-position curve indicated that memory is better for stimuli presented at the beginning of the sequence, called primacy effect and stimuli presented at the end of the sequence, called recency effect. The explanation offered for primacy effect is that stimuli presented first receive maximum attention and rehearsal so that they are encoded into the long-term memory while the explanation for recency effect is that stimuli presented last are still in the short-term memory.

Meanwhile, the original notion of short-term memory was a unitary store that serves only to hold information temporarily (Atkinson & Shiffrin, 1968). However, Baddeley and Hitch (1974) re-conceptualized short-term memory as a four-component, working memory system that includes both storage systems and a medium for processing of information. The processing component of the working memory was called the central executive and was said to be involved whenever humans engage in complex cognitive tasks e.g. solving a problem. According to Baddeley and Hitch's perspective, the working memory works closely with the long-term memory whenever humans engage in complex cognitive tasks. By extension one can infer that the working memory interacts with the long-term memory when participants engage in standard setting tasks. With the latter inference implies that it is appropriate to review also the long-term memory system.

Most reference to memory is to the long-term memory. The long-term memory was defined as a permanent archive of information about our past experiences (Goldstein, 2008). Even though there is still controversy about the number of the long-term memory systems, however given the variety of information it handles the notion that there is only one long-term memory store is highly improbable (Eysenck & Keane, 2010). It is generally accepted that there are two major types of long-term memory system namely: implicit and explicit memory (Gardiner & Java, 1993; Graf & Schacter, 1985; Schacter, 1990; Schacter & Tulving, 1994;

Tulving, 1983, 1985a, 1985b; Tulving & Schacter, 1990). These long-term memory systems were also referred to as declarative and non-declarative, respectively (Eysenck & Keane, 2010). The declarative and non-declarative distinction of the long-term memory systems is more intuitive and was adopted for the rest of this review.

The declarative memory denotes memories that can be described and that involve conscious recollections of previous experiences (e.g. events we have experienced or facts we have learned), while non-declarative memory denotes influence of past experience and enhanced performance in the absence of conscious recollection (Eysenck & Keane, 2010; Goldstein, 2008; Roediger, 1990; Schacter, 1987; Tulving, 1985b). Declarative and non-declarative long-term memory systems are distinguished in terms level of consciousness involved when using the information stored in them (Jacoby, 1983, 1988; Jacoby, Kelly, & Dywan, 1989). Precisely, we humans are aware when using the declarative memory, but unaware of use of the non-declarative memory. There are two types of declarative memory systems namely, the episodic and the semantic. Likewise, there seemed to be consensus on two major types of non-declarative namely: perceptual representation system and procedural memory (Eysenck & Keane, 2010; Goldstein, 2008). However, because this dissertation is about deliberate cognitive processes, this review is restricted to the declarative types of memory. Interested reader can reference the chapters on memory by **(**Eysenck & Keane, 2010; Goldstein, 2008) for details about the non-declarative types of memory.

Now back to the declarative memory. There are two main types: the episodic and the semantic. Episodic memory is memory for personal events with associated contextual information such as when and where they took place (e.g. remembering running into a friend at the library yesterday afternoon while semantic memory is memory for general knowledge such as facts (Eysenck & Keane, 2010; Goldstein, 2008). There is also convincing evidence based on study of amnesic patients that these two systems are different (Eysenck & Keane, 2010). As indicated at beginning of this section, this review of memory systems is restricted to concepts so that it assumed based on empirical evidence that they exist. Interested reader can check out Spiers, Maguire, and Burgess (2001) for empirical evidence pertaining to the episodic and the semantic memory.

The episodic and semantic systems are distinguished in terms of types of information they retain and the experience associated with retrieving information from them (Goldstein, 2008). While the episodic memory retains events and retrieval of information from it requires mental time travel in order to reconnect with the past as it was experienced, in contrast semantic memory retains knowledge that is not necessarily tied to specific personal experiences and therefore retrieval does not involve mental time travel (Goldstein, 2008; Tulving, 1985b). On the other hand, semantic and episodic memory are related because semantic memory can be conceived of as episodic memory that faded away leaving only the gist of the experience (Goldstein, 2008).

The types of knowledge stored in the semantic memory is quite varied, including for instance information about the rules of a game, names of cities, etc. Most of what is known about the organization of information in the semantic memory is about concepts. Concepts are an essential form of knowledge and are mental representations that are used for a variety of cognitive functions including pattern recognition. In particular, the most often studied function of concepts is categorization which is the process by which things are placed into groups called categories (Goldstein, 2008). A major application of concepts stored in the semantic memory is in the cognitive process of categorization (Eysenck & Keane, 2010). Therefore, the review of this literature is deferred to the section on categorization. In the meantime, a snap shot review of long-term memory processes is provided in the ensuing discussions. There are three long-term memory processes namely: encoding, storing, and retrieving (Berstein & Nash, 1999; Eysenck, 2010; Goldstein, 2008).

Encoding denotes processes occurring during the presentation of stimulus material that enables getting information from the working memory into the long-term memory. Encoding is facilitated by Rehearsal, a process that controls movement of information from the working memory into the long-term memory. According to level of processing theory (Craik & Lockhart, 1972), there are two types of rehearsals: maintenance and elaborative. Elaborative rehearsal constitutes deeper processing of information that involves attention to meaning and relating new information to those already stored in long-term memory. Maintenance rehearsal involves shallow processing of information such as simple repetition of information and attention to surface features of the stimulus material (Goldstein, 2008). Elaborative rehearsal is more effective than maintenance for encoding information into the long-term memory. Elaborative rehearsal facilitates deeper levels of analyses which produces more elaborate, longer lasting, stronger memory traces, and determines to a great extent how much we remember information over a long term.

Proper encoding of information leads to its storage in the long-term memory system. Retrieval then is the process for getting information from the long-term memory back into the working memory system. The accounts thus far suggest that remembering and retrieval process relies on learning and storing information. Therefore there cannot be retrieval without previous encoding and storage of information. Also, according to the information processing model (Atkinson & Shiffrin, 1968), for information to be firmly implanted in memory, it must pass through the three stages of mental processing (i.e. sensory, short-term, and long-term processing).

There is ample research evidence on the relationship between the manner in which information is encoded and the capacity to retain and retrieve it when needed. Reviewed next are some of these types of evidence, precisely about the relationship between encoding mechanism and performance on memory retrieval. Retrieval can be enhanced by organizing information and with the use of cues (Goldstein, 2008). There is evidence, that organizing material results in substantially better recall (Bower, Clark, Lesgold, & Winzenz, 1969). One suggested way of organizing information for recall is in terms of categories, in particular it was noted that remembering words in a particular category can serve as retrieval cues for words in that same category (Bransford & Johnson, 1972; Goldstein, 2008; Jenkins & Russell, 1952).

Moreover, semantic coding of information in the LTM is also noted as particularly important for retrieval. For instance, there is evidence based on levels of processing research paradigm that maintenance rehearsal (e.g. repetition of stimulus information), typically has small effect on long-term memory (Glenberg, Smith, & Green, 1977). Based on this evidence the conclusion was that rehearsal was necessary but not sufficient for proper encoding (Eysenck & Keane, 2010). Evidence from levels of processing research Craik and Lockhart (1972), using incidental learning, whereby the participants do not know that there would be a memory test at the time of learning, also showed the following: that the greater the extent to which the meaning of information is processed during learning the better the gist of it can be remembered (Craik & Tulving, 1975; Sachs, 1967); the greater the amount of processing of a stimulus during learning the better the memory for it, this confirmed the elaboration of processing hypothesis (Craik & Tulving, 1975); distinct memory traces are easier to retrieve than those similar to other memory traces, this confirmed the distinctiveness hypothesis (Eysenck, 1979; Eysenck & Eysenck, 1980).

Likewise, evidence in favor of cues is demonstrated in the laboratory using cued recall and is also provided by experiments that show encoding specificity principles **(**Tulving, 1979) that posit that information is learned together with its context or state dependency principle that posit that information is learned in association to a particular internal state. A classic experiment that demonstrates encoding specificity showed that best recall occurred when encoding and retrieval occurred in the same location (Godden & Baddeley, 1975). The power of self-generated retrieval cues was also demonstrated in an experiment that showed near perfect recall for participants when they were given cues they created (Mantyla, 1986). Another reported strategy that can help to enhance retrieving information from the long-term memory is to tailor encoding process to the task required at the retrieval end. Evidence for the latter is based on the transfer appropriate processing theory of memory (Anderson, 1990; Baddeley, 1992), which posits that recall is enhanced if task at encoding matches task at retrieval, regardless of level of processing e.g. storing semantic information is irrelevant when the memory test requires the identification of capitalized words, this confirmed the transfer appropriate processing hypothesis (Morris, Bransford, & Franks, 1977).

It is important to also review measures of memory. There are two main tests for episodic memory namely: recall and recognition. Both recall and recognition refer also to memory tasks. The essence of recall is to generate information meeting the definition of the target in the recall instruction. On the other hand, in recognition tasks, one or more potential targets are presented to subjects with no overt instruction of generation. According to the two-process theory, while recall involves both search and decision, recognition involves only decision so that recognition task is one of discrimination while the recall task is one of retrieval (Brown, 1976; Kintsch, 1970).

The most basic recognition memory test involves presenting a series of items, with participants deciding whether each one was presented previously. Recognition memory was said to be mediated either by recollection or familiarity (e.g. Mandler, 1980). According to Eysenck and Keane (2010), citing Diana, Yonelinas, and Ranganath (2007), while recollection is the process whereby we recognize a learned item with contextual details, familiarity is the process of recognizing an item without any specific details of the learning episode (e.g. we might recognize someone's face as familiar without precisely remembering where and when we came across the person).

There are three basic forms of recall test: free recall, serial recall, cued recall. First, free recall involves generating learned material in any order and in the absence of cue. Second, serial recall involves producing learned material in the order in which it was presented. Third, cued recall involves producing learned material in the presence of cues e.g. one may have learned cat-table and in the test the cue cat- might be tested (Eysenck & Keane, 2010, p. 260). The supposition is that the Angoff standard setting method task which is the focus of this dissertation necessarily involves free recall test. Consequently, the remaining review of measures of memory is restricted to research on recall.

Research on recall focused on whether the processes underlying recall were the same as those involved in recognition memory. Eysenck and Keane (2010) reported an example of such study conducted by Staresina and Davachi (2006) in which they used three memory tests namely: free recall, item recognition (familiarity), and associative recognition (recollection). The finding from this study was that successful performance of the tasks was related with greater involvement of brain areas associated with the episodic memory during encoding. The researchers concluded that successful free recall involves forming associations which is not required for successful recognition memory. Eysenck and Keane (2010) arrived at the following conclusions based on these results namely: (1) the finding that similar brain areas are involved in free recall and recognition suggests that there are important similarities between the two types of memory test; (2) since successful free recall is associated with higher levels of brain activity in several areas at encoding and retrieval than successful recognition it indicates that free recall is in some sense more difficult than recognition memory; and, (3) the finding by Staresina and Davachi's (2006) study that some areas of the brain associated with successful free recall are not involved in recognition memory suggests that free recall involves additional processes than those involved in recognition memory e.g. inter-item processing is specific to free recall. The last line of memory research reviewed in the immediately following paragraph is that investigating the accuracy of episodic memory.

Another important line of research on memory that can inform standard setting research focused on whether episodic memory was reproductive i.e. provides accurate and detailed information about past events or constructive i.e. prone to various kinds of errors and illusions (e.g., Schacter & Addis, 2007). There is sufficient evidence that episodic memory is constructive (Druckman & Bjork, 1994). What people report as memories are constructed by the person based on what actually happened plus additional factors, such as the person's knowledge, experiences, and expectations. The latter implicates that the mind constructs memories based on a number of sources of information and that errors in memory emanates from the influence of our generalized knowledge (Harris, Sardarpoor-Bascom, & Meyer, 1989). For instance, the constructive nature of the episodic memory leads the eye witnesses to report distorted memories of what they had seen (Eysenck & Keane, 2010). Also, work by Bartlett (1932) asserted that the knowledge we possess can lead to systematic errors in our episodic memories and this assertion was confirmed by empirical research. Three explanations were offered for the constructive nature of the episodic memory namely: (1) it would require a great deal of processing to produce semi-permanent record of all our experiences; (2) we generally assess the gist of our past experiences not trivialities; and, (3) the constructive processes are also used for imagining future events (Eysenck & Keane, 2010; Schacter & Addis, 2007).

Meanwhile, the hypothesis that we humans typically remember the gist of what we experienced and that this tendency increases with age have been tested empirically (an example was study by Brainerd & Mojardin, 1998). The result from the latter study showed that false recognition increased with age. Also tested empirically was the hypothesis that imagining future event involves the same processes as those involved in remembering past experiences (e.g., Hassabis, Kumaran, Van, & Maguire, 2007). Hassabis, Kumaran, Van, and Maguire (2007) presented evidence that amnesic patients produced imaginary experiences with less richness and spatial coherence than the healthy group. In addition this hypothesis was also investigated by comparing brain activity when individuals generated past and future events and elaborated on them (Addis, Wong, & Schacter, 2007). The finding was that there was higher activity in several areas of the brain during generation of future events than of past events which was interpreted as indicating that more constructive processes are required for imagining future events than to retrieve past events (Eysenck & Keane, 2010). The overall conclusion from this line of research was that episodic memory was constructive rather than productive. Also, studies based on repeated recall tests suggest that memory for an experienced event becomes more constructive over time (Brown & Kulik, 1977).

To summarize, the take away from reviewed memory research for this dissertation study were as follows: (1) working memory and long-term declarative memory systems (i.e. semantic and episodic memory) are most probably engaged in standard setting judgment tasks so that participants should be recruited with sound working memory and declarative memory of interaction of students with test items; (2) the Angoff task necessarily involves free recall of information relevant to the judgment task; (3) the levels of processing principles (e.g., deep semantic processing, elaborative processing, use of cues, matching tasks at encoding and retrieval, etc.) should serve as guide for design of training instructions, practice, and feedback to enhance participants recall performance; (4) training instructions should emphasize retrieval of past events as opposed to imagination of present or future events to increase veridicality of memory reports; (5) since the availability heuristic, which entails selective recall of information (precisely recall of what comes to mind first) is reported to lead to error, participants should be encouraged to engage in extensive search of memory and, (6) participants should be selected that have recent experience with the target population of students for which performance information is elicited to reduce errors in memory reports.

B. Categorization: Concepts and Theories

As mentioned in the review section of the semantic memory, most of what is known about the organization of semantic memory is about concepts**.** Concepts were defined as the mental representations corresponding to categories of objects or items in the world (Murphy, 2002; Eysenck & Keane, 2010). Two lines of research on concept would be reviewed first in the immediately following paragraphs before the process of categorization which necessarily depends on concepts.

The first line of research on concepts focused on how concepts are organized in the semantic memory. The assumption was that concepts were organized in a hierarchy. By hierarchical organization concepts are interconnected in a tree form beginning with the more general e.g. furniture (at the top) down to specific e.g. high chair (at the bottom) (Goldstein, 2008). This approach to organization of concept was called the semantic network approach **(**Collins & Quillian, 1969; Quillian, 1967, 1969). The limitation of the semantic network model was that it assumes that concepts belong to rigidly defined categories this was said to be mistaken as there is convincing evidence that many concepts in the semantic memory are fuzzy (Eysenck & Keane, 2010; McCloskey & Glucksberg, 1978). Collins and Loftus (1975) put forward spreading activation theory and argued that the notion of logically organized hierarchies was too inflexible. They assumed instead that semantic memory was organized in terms of semantic relatedness or semantic distance. Semantic relatedness is measured by asking people for instance, to decide how closely related pairs of words are or to list as many members as they can of a particular category. The assumption of the spreading activation theory was that when a person sees, hears or thinks about a concept, the appropriate node in hierarchy is activated and this activation then spreads most strongly to other concepts closely related semantically, and more weakly to those more distant semantically (Eysenck & Keane, 2010).

The second line of inquiry on concepts sought to find out if there is a basic psychologically privileged level of concepts (Goldstein, 2008). Rosch, Mervis, Gray, Johnson, & Boyes-Braem (1976) research started with the notion that with the hierarchical model of organization of concepts, there are three different levels of concepts, ranging from the most general to specific and that when people use concepts they tend to focus on one of these levels. She distinguished three levels of concepts as follows: superordinate level (e.g. furniture), the basic level (e.g. table), and the subordinate level (e.g. kitchen table). She proposed that the basic level is psychologically special because it is the level above which much information is lost and below which little information is gained.

Now to the process of categorization, categorization is a cognitive process that relies on stored concepts in the semantic memory. By categorizing objects are placed into groups called categories (Goldstein, 2008). There are two theories about the mental processes underlying categorization namely: similarity judgment and theory. Citing Murphy and Medin (1985), Murphy (2002) indicated that theory based categorization approach suggests that categorization is not simply based on direct matching of properties of the concepts with those of an example but rather requires that the example have the right explanatory relationship to the theory organizing concept. The theory based categorization approach thus views the relation between a concept and an example as analogous to the relation between theory and data. It was deemed appropriate to make reference to these two theories to highlight that categorization is not always accomplished through the process of similarity judgment between an object under consideration and the mental representation of categories. However, the preferred view of categorization in this dissertation is a process that is mediated by similarity judgment which fits well with the representativeness and the availability judgment heuristics framework of this dissertation. Therefore, this review would focus on theories of categorization that are related to the notion of similarity judgment.

There are three theories about the structure of categories or about approaches to categorizing objects that rely on similarity judgment: definition, prototype, and exemplar. These three theories assume that similarity is the critical category-forming relation (Sloman & Rips, 1998). By the definitional approach to categorization an object is designated as a member of a category, if it meets the definition of the category. Smith and Medin (1981) called the definitional approach the classical view in order to distinguish it from the modern views of categorization which are based on the notion of probability. In order to maintain consistency with the latter distinction, classical view would be used for subsequent discussions. According to Murphy (2002), there are three major claims of the classical theory of categorization:(1) concepts are mentally represented as definitions which provides characteristics that are necessary and sufficient for determining membership of objects to the category; (2) An object is either in or not in a category with no in-between cases; and, (3) It does not make distinction between category members. The classical view makes restrictive assumptions about conceptual structure and therefore, fit only well-defined categories. Precisely, the definitional approach was said to work well for geometrical objects, e.g. the category of triangle, but not for natural and human objects (Goldstein, 2008). As a consequence, the classical theory was rejected and with this revolution came the ascendancy of probabilistic notions of conceptual structure.

The probabilistic notions hold that categories are fuzzy or ill-defined so that they are organized around a set of properties typical of category members (Murphy, 2002). There are two probabilistic views of conceptual structure namely: exemplar and prototype. These probabilistic views posit that concepts are organized around the notion of family resemblance (Rosch & Mervis, 1975; Wittgenstein, 1953). Wittgenstein (1953) introduced the idea of family resemblance, to denote the fact that objects in a particular category resemble one another in a number of ways thereby allowing for variation within a category. Meanwhile, within category variation represents differences in prototypicality (Rosch, 1973). While high prototypicality means that the category members closely resemble the category prototype, low prototypicality implies little overlap between members of the category. It is this idea of family resemblance that underlies the prototype and exemplar approaches to categorization. Precisely, the prototype and exemplar approaches to categorization are both based on the idea that membership in a category can be determined by comparing an object to a standard that represents the category.

The prototype view of conceptual structure assumes that categories are organized around an ideal summary representation or an example that represents all of the characteristic features of the category. Put differently, the assumption of the prototype approach is that that we abstract the central tendency of examples of a category as a mental representation for the category. Therefore, according to the prototypical view, categorization proceeds by judgment of similarity between an object with the prototype of a category. In contrast to the prototype view, the exemplar view assumes that there is no single summary representation rather mental representation of a category consists of a set of encoded examples of the category. Although the prototype and exemplar views are both based on the similarity principle of categorization, they differ in terms of supposition of the standard to which an instance is compared in order to be classified. Precisely, while the prototype view assumes that the mental representation is a prototype expressed as the average of members of a category, the exemplar view assumes that it is a set of encoded examples. On the other hand, the exemplar approach does better than the prototype approach because besides central tendency, it incorporates other category information such as category size, variability of examples, and information concerning correlation between attributes (Murphy, 2002).

So far, the concept of similarity has being used simply in terms of a construct mediating the cognitive process of categorization. However, because of the compelling explanatory role that the construct of similarity plays in theories of categorization (i.e. the classical, the exemplar, and the prototype theories), it is appropriate to also to explicate the theories of similarity judgment. Consequently, the rest of the review in this section is aimed at explicating on the fundamental mental operations involved in similarity judgment. There are two theories about similarity judgment process also referred to as similarity representation theories: the geometric and feature matching theory of similarity.

The geometric view of similarity assumes that the judgment of similarity between objects can be adequately modeled as the computation of metric distance between points. The term "object" as used encompasses persons and stimuli. The metric distance function is a scale that assigns a non-negative number to every pair of points called their distance in accordance with three axioms namely: minimality, symmetry, and triangular inequality. The minimality axiom assumes that the similarity between an object and itself is always greater than the similarity between it and other objects, and is unity; the symmetric axiom posits that the similarity between two objects is the same regardless of the direction in which the similarity relation is expressed; while the triangular inequality assumes that similarity relation is transitive so that if object one is similar to object two and object two is similar to object three implies that the object three is similar to object one. Geometric models represent objects in a coordinate space so that the observed dissimilarities between objects correspond to the metric distances between the objects.

In making a case for his feature theory of similarity, Tversky (1977) highlighted that even though the geometric models of similarity have achieved some success in empirical applications to similarity relations (see for example, Carroll & Wish, 1974; Nosofsky, 1988a, 1988b; Shepard, 1974; Smith, Shoben & Ribs, 1974; for multidimensional scaling), that the geometric approach faces several difficulties. His argument against the geometric approach was that the applicability of the dimensional assumption was limited for analysis of similarity between certain stimuli and that the metric axioms are questionable. For instance, the minimality axiom is problematic, symmetry is false, and the triangle inequality is not compelling. Therefore, he surmised that it seems more appropriate to represent some stimuli (e.g. personalities), in terms of many qualitative features than in terms of a few quantitative dimensions. Therefore the assessment of similarity between such stimuli may be better described as a comparison of features rather than as the computation of metric distance between points.

Consequently, the feature matching theory Tversky's (1977) was the proposed alternative theoretical approach to similarity judgment. The feature matching theory as described in this paragraph is mostly a summary of the original conceptual paper on the approach (Tversky, 1977). The feature matching theory was formulated in terms of the set theoretical notion of matching function rather than in terms of the geometric concept of distance. Therefore it is neither dimensional nor metric in nature (Gati & Tversky, 1984; Tversky, 1977). The term feature as appears in the label of the theory denotes the value of a binary variable (e.g., gender) or the value of a nominal variable (e.g., eye color). Feature representations, however, are not restricted to binary or nominal variables; they are also applicable to ordinal or cardinal variables (i.e., dimensions). With the feature matching approach, objects are represented as a collection of features and similarity is described as a feature matching process.

In formal terms, the feature matching theory assumes that the similarity between two objects, say *a* and *b*, is expressed as a matching function of three arguments namely: the features that are common to both objects, the features that belong to object *a* but not to object *b*, and the features that belong to object *b* but not to object *a*. In addition, the matching function assumes monotonicity so that the similarity between object *a* and object *b* exceeds similarity between object *a* and object *c*, if object *a* and object *b* have more common feature and fewer distinctive features than object *a* and object *c*. The aforementioned couple of assumptions are called the matching function because together they measure the degree to which two objects match each other. However in order to determine the functional form of the matching function, additional assumptions about similarity are introduced namely: independence, solvability, and invariance. Solvability requires that the feature space under study be sufficiently rich so that certain similarity equations can be solved while invariance ensures that the equivalence of intervals is preserved across factors.

Meanwhile, the model of the feature matching theory of similarity presented in Tversky's (1977) paper was the contrast model. The contrast model was based on the overarching idea that our total data base concerning a particular object (e.g., a person) is generally rich in content and complex in form that includes appearance, function, relation to other objects, and its properties that can be deduced from our general knowledge of the world. As a consequence, when we are faced with a particular task (e.g., identification or similarity assessment) we extract and compile from our data base a limited list of relevant features on the basis of which we perform the required task. Thus, the representation of an object as a collection of features is viewed as a product of a prior process of extraction and compilation. In formal terms, the contrast model expresses the similarity between objects as a linear combination, or a contrast, of the measures of their common and distinctive features. The major constructs of the contrast model were the contrast rule for the assessment of similarity and scale which reflects the salience or prominence of the various features. The scale denoted *f* measures the contribution of common and distinctive features to the similarity between objects. With contrast model formulation necessarily follows that increase in the measure of the common features increases similarity and decreases difference, whereas an increase in the measure of the distinctive features decreases similarity and increases difference between objects.

To conclude this review of theories related to categorization, it is appropriate to recap that categorization is the fundamental principle of the representativeness heuristic. The assumption in this dissertation based on the representativeness heuristic is that judgment of conditional probability is partly mediated by categorization. Meanwhile, categorization is mediated by similarity judgment. Precisely, categorization is based on determination of the similarity between the objects with prototypes or examples of the experienced instances of the categories to which the object might belong. Similarity judgment between an object and a category is assumed to be based on categorical or ordinal feature comparison process that involves feature extraction and integration. Therefore the appropriate model for similarity judgment is the feature matching model.

The feature matching theory of similarity judgment substantively makes more intuitive sense as a model of cognition for the standard setting participants. In formal terms, the feature model extends object features basis for similarity judgment to include qualitative and ordinal variables and it generalizes standard representations of similarity data to include clusters and trees. It is important to highlight that there is a correspondence between features of objects and the classes to which the objects belong, which provides a direct link between feature matching theory and tree and clustering approach to the representation of proximity data (Tversky, 1977).

Finally, the main take away from the feature theory for the task of training was that similarity relation is incomplete without specification of the respects for similarity judgment and the rules for weighting common and distinctive features Medin, Goldstone, and Gentner (1993). Therefore, the prescription of rules for categorization should explicate the features for judgment and emphasize paying more attention to common features than distinctive of objects while judging their similarities. The latter process ensures that the relative weight of common features is greater when making judgment of similarity between objects.

2.4. Theoretical Framework for Social Psychology

This section summarizes the key social psychological findings of the effect of group discussion on a groups judgment and decision making process as reported by Fitzpatrick (1989). The most reported influence of social interaction on group judgments which has great generality is called Group-Induced Polarization. Group-Induced Polarization is the shift of a group's judgment to a more extreme position following discussion**.** This concept is operationalized in terms of shift of mean to a more extreme left or right of the mean. This observed effect of social interaction on group judgment in terms of mean comparison contradicts the commonly held notion that social interaction among group members involves conformity pressures that simply produce convergence on the central tendency of the group. However, in those studies where pre- and post-discussion variability was compared, a convergence of judgment was typically observed to follow group discussion (Lamm & Myers, 1978; Stoner, 1961). Two theories advanced to explain the polarizing effect of social interaction have gained the most empirical support and were:

- The theory of social comparison - Polarization is the product of interpersonal comparison processes in which group members compare their opinion to those held by similar others in other to ascertain their accuracy (Festinger, 1954). This theory posits that uncertainty of soundness of judgment is a key predictor of revision of opinion to make it more like the opinion held by the similar others. Hence, the higher the uncertainty involved in a judgment task, the more likely it becomes that the primary effect of group discussion is normative and that a shift in the mean post-discussion opinion of the group occurs.

- The theory of informational influence - According to this interpretation, group discussion generates arguments, cognitive learning occurs, and group members revise their judgments in the preferred direction and polarization is observed. This view posits that discussion provides a forum of exchange and generation of information that leads to cognitive learning, reduces uncertainty and leads to revisions of judgment in the expected direction of the issue.

The key conclusions about the findings of group-polarizations in social interaction studies were as follows: (1) the effect is more likely to be observed with subjective judgments than with objective ones; (2) both social comparison and informational influence are present in discussion settings; and, (3) polarization is inhibited under circumstances of exposure to the distribution instead of summary statistic of group members opinion positions, when group members are asked to record arguments supporting their opinions, are distracted from thinking about the opinions of the group members, and when they are publicly bound to their opinion. Given the conceptual views in this dissertation, social influence generated by interpersonal comparisons are undesirable while cognitive learning through the exchange of information is desirable. Precisely, influence through better cognitive learning of categorical information pertaining to the test knowledge and skill constructs in standard setting is desirable.

Chapter Three: Literature Review

This chapter reviews standard setting research literature with specific reference to the Angoff (1971) method. There are five major sections in this chapter. The first section presents a historical account of the Angoff method; the second section reviews evidence for and current training approaches to addressing cognitive complexity of Angoff method tasks; the third section briefly reviews alternative standard setting approach to addressing cognitive complexity of the Angoff method tasks; the fourth section reviews research on cognitive processes underlying the Angoff method tasks; and, the fifth and last section presents the motivation for the Heuristic training approach of this dissertation to addressing cognitive complexity of the Angoff method tasks. In all of these reviews, discussion of standard setting research is intermixed where necessary with conceptualizations from research in the related cognitive psychology field in order to draw out the theoretical underpinnings.

3.1. Historical Review of the Angoff Method

From a historical standpoint, the Angoff (1971) method was one of the first prescribed criterion- referenced, judgmental, and test-centered standard setting methods (Zieky, 2001). The Angoff method was named after William Angoff, who first proposed it in the year 1971, in a book chapter on test equating, scaling, and norming. However, according to Zieky (2001), William Angoff had indicated in a personal communication that he was never comfortable that the method was named after him and maintained that the idea for the method came from a colleague, Ledyard Tucker.

The Angoff method as originally formulated, what would be called an unmodified Angoff method, was described in the footnote of the book chapter and required judges to:

"State the probability that the "minimally acceptable person" would answer each item correctly. In effect, judges would think of a number of minimally acceptable persons, instead of only one such person, and would estimate the proportion of minimally acceptable persons who would answer each item correctly” (Angoff, 1971, p. 515).

As an aside, Angoff (1971) also prescribed in the text of the chapter contribution, a deterministic version of the unmodified Angoff method called the Yes/No method which required participants instead to simply indicate a “yes” response if the target group of students would respond correctly to an item and a “no” response if the target group of students would not respond correctly to an item. These “yes” and “no” responses are subsequently converted into “1” and “0”, respectively, and scores are summed for each participant and averaged across participants to derive the cut score. Although the stochastic unmodified Angoff method requiring instead, judgment of probability of a correct response for each test item, was described in the footnote, it has since remained the more popular method (Cizek, 2001; Cizek & Bunch, 2007; Impara & Plake, 1997). Meanwhile, the same process of summing probability estimates for each participant and averaging across participants is most often followed in order to derive cut score for the unmodified Angoff method. Angoff did not provide rationale for the Yes/No or for the unmodified Angoff method (Impara & Plake, 1997). However, it seems that standard setting researchers recognized the statistical foundation of both methods. This is evidenced by preference for the unmodified Angoff method because, although it posits more complex tasks for participants, it is an improved conceptualization of test taking episodes, as generators of outcomes that are chance events as opposed to deterministic events.

The unmodified Angoff method as conceived was based on frequentist statistical parameter estimation principles which have been established to be robust and to yield unbiased parameter estimates under circumstances that assumptions are reasonably well met. The method involves two tasks namely: (1) conceptualizing the student population i.e. the minimally acceptable persons and, (2) estimating the proportion of the minimally acceptable persons who would answer each of dichotomously scored items correctly (Hein & Skaggs, 2009, 2010; Raymond & Reid, 2001).

As originally formulated, the first component task of the method, of conceptualizing the target student group was the more explicated in operational terms and has remained the more researched. For instance, the original prescription indicated that a group of minimally acceptable persons should be conceptualized and that the probability metric should be operationalized by estimating the proportion of the minimally acceptable persons who would answer each item correctly. These item proportion correct estimates in the educational measurement context are item difficulties for the minimally acceptable persons. On the other hand, the original formulation did not explicate how to generate the item difficulties on the proportion correct scale. Hence, current training for the second component task emphasize introduction of feedback modifications and without reference to its knowledge and skills requirements.

This state of affairs of under explicating the second component task of the Angoff method operationally still persists. The assumption has been that the participants know how to estimate item difficulties and are able to do so accurately (Wyse, 2009). Therefore, this dissertation relaxes the latter assumption by explicating further the Angoff method and through training instructions on what factors to consider and on how to estimate item difficulties. Meanwhile, the conventional naming scheme of referring to operational explications of the unmodified Angoff method as “modified Angoff method” (Cizek, 2001; Cizek & Bunch, 2007; Zieky, 2001) would be adopted by this dissertation.

Nevertheless, since its introduction, the Angoff method has remained one of the most widely applied and the most researched method (Cizek, 2001; Cizek & Bunch, 2007; Hein & Skaggs, 2010; Skaggs & Hein, 2011; Wyse, 2009; Zieky, 2001). Penchant for the Angoff method may be because of its strong statistical and psychometric theoretical underpinnings. At the same time, it holds the record of been the most critiqued and attacked by researchers (see for example, Glass, 1978; Schultz, 2006; Shepard, 1995; Shepard et al., 1993). It was not long after the Angoff method was introduced when attacks began to be launched on the method. Much of the empirical research work on the method within the first three decades of its introduction that spurred these criticisms is reviewed in Brandon (2004). The fundamental critique of the Angoff method is that its tasks are cognitively complex for execution by participants (Hein & Skaggs, 2010; Impara & Plake, 1997, 1998).

A couple of widely cited attacks were launched within the first two decades of researching the method. The first major attack was launched by Glass (1978) in the *Journal of Educational Measurement* who concluded that participants judgments was substantially unreliable so that setting performance standards on tests by known methods was a waste of time (Skaggs & Hein, 2011; Wyse, 2009; Zieky, 2001). Subsequently, after roughly two decades of research, the second attack was launched by Shepard et al. (1993). Zieky (2001), quoting Shepard et al. (1993) described the Angoff method and other methods based on the judgments of items as "fundamentally flawed" and suggested that even improvements cannot neutralize the nearly impossible cognitive task requirements. Within the last decade additional criticism of the Angoff method by Schultz (2006) was that participants' judgments suffers from regression to the mean of the probability scale (Wyse, 2009).

These attacks on the Angoff method culminated and caused a switch to use of the Bookmark procedure, a presumably simpler method. The Bookmark method has now become the method of choice in the public school standard setting contexts (Karontonis & Sireci, 2006). At the same time, although a number of standard setting researchers have called for adequate emphasis of training, it still remains one of more underemphasized (Clauser, Swanson , & Harik, 2002; Cross, Impara, Frary, & Jaeger, 1984; Mills, Melican & Ahluwalia, 1991; Plake, Melican, & Mills, 1991; Raymond & Reid, 2001; Reckase, 2000, 2001; Reid, 1991). To underscore the little emphasis on training, Brandon (2004) in his review of research on the Angoff method identified nine most commonly studied topics, none of which addressed cognitive processes of participants or training.

Therefore, the argument made in this dissertation is that the Angoff method holds prospects for standard setting research if only research fully explicates and explores the tasks. Although it might be a leap of inference, the main claim of this dissertation is that adequate explication of the Angoff tasks in the design of training and appropriate evaluation of these training procedures would lead to desired state of affairs of veridicality of outcomes of laboratory Angoff standard setting. In the remaining sections of this chapter, previous research efforts at addressing the cognitive complexity of the Angoff method tasks and their shortfalls are reviewed and the motivation for the Heuristic training approach proffered by this dissertation is presented.

3.2. Research on the Angoff Method Training

It is appropriate to begin this section review with the assumptions of the original formulation of the Angoff method and the evidence for cognitive complexity of its tasks. The assumptions were that the participants are able to (1) conceptualize the minimally acceptable student group by identifying the knowledge, skills, and ability levels typical of this group of students; and, (2) predict how well the students would perform on each item on a test on the proportion correct metric (Impara & Plake, 1998). However, past and current research evidence suggests that both tasks are cognitively complex for Angoff method participants.

For instance, evidence based on introspective reports of participants suggests that teachers are better able to think in terms of specific and real students in their classroom than about a hypothetical group of students (Hein & Skaggs, 2010; Impara & Plake, 1997; Skaggs & Hein, 2011). Also, according to Impara and Plake (1997) the idea of using a panel of participants to estimate the proportion of a group of students who would correctly answer an item (or item difficulty) was not new in 1971. It had been attempted in the past by Lorge and Kruglov (1953) who had reported that participants were unable to estimate item difficulty accurately even though their rank ordering of items in terms of difficulty were modestly accurate[2]. In other words, correlations between participants estimated item difficulties and the actual item difficulties were positive and moderate but the estimates were obviously discrepant from the actual values.

This finding of inability of standard setting participants to estimate absolute item difficulty has been replicated in several subsequent studies and consequently, was spotted as a problem (Bejar, 1983; Halpin & Halpin, 1983; Schaeffer & Collins, 1984; Shepard, 1994; Thorndike, 1980). Evidence from several standard setting studies demonstrated that participants are not universally competent in estimating absolute item difficulty (Impara & Plake, 1997, 1998; Reid, 1991). For instance, citing Shepard (1994), Impara and Plake (1998) reported that trained judges systematically erred in their estimates of item performance by overestimating

[2] Please note that item difficulty judgment is the substantive reference for the Angoff method conditional probability judgment task in educational measurement field.

examinee performance on difficult items and underestimating examinee performance on easy items. She concluded based on this finding that the Angoff method may not provide valid cut scores because the participants are unable to perform the major task required of them which is estimating the conditional probabilities of correct response for test items.

Inability of standard setting participants to estimate item difficulty was reported even when participants were familiar with the students and with the test. In their study, Impara and Plake (1998) tested the ability of classroom teachers to estimate item performance for two groups of their students on a locally developed district-wide science test. The findings were that teachers were more accurate in estimating the performance of the total group than of the "borderline group”, but in neither case was their accuracy level high. Hence, in their report of the findings Impara and Plake indicated that if participants were unable to make accurate predictions even for the ideal case that knowledge assumptions were met that it calls into question whether the tasks posed by the Angoff method were realistic and thus the validity of Angoff method assumptions.

Their conclusion likewise that of Impara and Plake (1997) was that little confidence may be placed on the accuracy of participants estimated proportion correct even when the participants have high degree of familiarity with both the examinees and the test. Therefore, their projection was that the participants will likely have an even harder time estimating proportion correct accurately for items they may have never seen before and for a group of hypothetical examinees representing real examinees they have little or no experience with (Impara & Plake, 1997). The latter projection was notable because contemporary Angoff studies often recruit participants who do not have the advantage of possessing high levels of knowledge of the student group or much prior exposure to the test.

In addition to the reviewed standard setting research evidence, studies outside the area of standard setting provide evidence that the task of estimating item difficulty is cognitively challenging for judges (Bejar, 1983; Lorge & Kruglov, 1953; Thorndike, 1980). In the latter studies, the general findings also were that judges are able to rank order items accurately in terms of difficulty but they are not particularly accurate at judging the absolute difficulty (i.e., the percentage of a reference group that will answer correctly (Lorge & Kruglov, 1953). As a consequence of the evidence of cognitive complexity of the Angoff method tasks, a number of conceptual articles have made recommendations on how to design training instructions to assist participants with the task of estimating item difficulties. A few of the widely cited recommendations are reviewed as follows.

Earlier, Reid (1991) had identified little structured Angoff method training as a gap in the standard setting research literature and indicated that it was unclear whether this was due to deliberate decision based on conviction that training participants was unnecessary or due to oversight. By structured training, Reid (1991) meant training that includes procedures to assist participants to arrive at a conceptualization of the target student group that barely makes in a performance category and in applying this conceptualization at the individual item level. He also highlighted the lacking evidence about the effect of pre-operational training activities on the item difficulty judgment task on participant's behavior. By pre-operational training he meant training instructions and practice activities prior to item performance data feedback rounds. The recommendation was that documenting such evidence was essential and he suggested that in the absence of such information that it seems reasonable to adopt a conservative approach and to assume that the lack of pre-operational training will have negative impact on the standard-setting exercise.

The article dwelt on prescriptions on how to train participants to generate item difficulty ratings once the reference group has been defined. With respect to training participants to generate item difficulty ratings, the recommendation was that because of the novelty of the task that it seems prudent that training should sensitize participants on factors that influence item difficulty. Although the suggestion that training should sensitize participants on factors that influence item difficulty was apt, however, the perceived shortcoming of Reid's proposal was with the advocated type of item difficulty judgment training. Precisely, the principles of the advocated type of training were consistent with the unbounded notion of human rationality[3]. To support the latter theoretical assertion is the following excerpts from the article about his recommendation:

> It is essential that judges understand that individual items are fallible measures of content. A judge who rates an item based upon what it appears to measure without tempering the rating based upon other factors impacting performance may under- or over-estimate item difficulty, depending upon the direction in which those factors act. Judges should be sensitized to factors within items which make them imperfect measures of the content that they are intended to measure. ….Beyond the general concept of the fallibility of items, training might include practice in evaluating items to sensitize judges to the factors other than content that affect an examinee's ability to answer correctly. Other factors such as item format, clarity of expression, and the "cognitive closeness" of distractors to the key may also have a significant impact on an examinee's ability to answer correctly. Alerting judges to these findings may help to improve judges' sensitivity to item difficulty (Reid, 1991, p. 12).

It is important to highlight that the basis for the recommendation inculcated in this excerpt was the assumption that insufficient information was the inherent problem with participant's inaccuracy in performing the task of estimating item difficulties. As a consequence,

[3] Please refer to the theoretical framework for cognitive psychology in Chapter two for description of this theoretical view

the proffered panacea to the problem of inaccuracy of judgment of item difficulty by participants was training on both unique and common features of items that determine item difficulty.

However, the shortfall of this recommendation is that it assumes limitless capacity of participants to process and integrate different types of information into their judgment. Besides, it overlooks the role of item pilot testing procedures which are supposed to take care of construct irrelevant factors that potentially impact item difficulty. Meanwhile, this view of item difficulty training is most probably shared by a good number of Angoff method researchers. The assumption of limitless human information processing capacity underlies the predominant approach to addressing complexity of the Angoff task through emphasis on individuating quantitative feedback information about items and devoid of processing instructions. This dissertation argues instead for human processing limitations as the factor delimiting utility of Angoff method.

More recently, Raymond and Reid (2001) reiterated the need for adequate emphasis on training instruction on the item difficulty judgment task. Their argument in favor of training instruction was that standard setting methods with specific reference to the Angoff method, involve tasks that participants are not likely to have performed previously. Moreover, although standard setting tasks may appear straightforward, discussions with participants during the course of standard-setting studies reveal a more complex cognitive process than what is evident on the surface. Therefore, it is unlikely that participants will understand or correctly perform the required tasks without benefit of training instructions.

Also, according to Clauser (2002), suggestions have been made by some standard setting researchers, that in addition to careful orientation to the procedure, and detailed discussion of the definition of the minimally competent examinee, that the training should include practice accompanied by feedback (Jaeger, 1989; Livingston & Zieky, 1982; Plake, Melican, & Mills, 1991; Reckase, 2000, 2001; Reid, 1991). The reviewed evidence and recommendations underscores the need for training for the Angoff method tasks albeit still unresolved in the literature is the type of training. In the immediately following subsection research on training instruction, practice, and feedback for the Angoff method tasks are reviewed.

3.2.1. Research on Training Instruction and Practice

Relatively more research studies have addressed through training instruction and practice, the task of conceptualizing a target group of students that barely makes it in a performance category (Giraud, Impara, & Plake, 2005; Hein & Skaggs, 2010; Impara & Plake, 1997). Besides the fact that it was the more explicated in the original prescription, the reason for the greater emphasis to this task is because it is shared by a number of other standard setting methods (Hein & Skaggs, 2010). For instance, Impara and Plake (1997) indicated that regardless of method used that the most perplexing problem encountered by participants of test-centered standard setting studies was how to conceptualize the target group of students.

Although the second task of the Angoff method of estimating item difficulties is equally important and arguably a more cognitively challenging task for participants, it has been given less attention in terms of training instructions and practice. Estimating item difficulty is a more cognitively complex task because, for a given student performance category, there are conceivably multiple substantive domain categorization of test items that have to be considered in order to generate meaningful item difficulties. The complexity of the task of estimating item difficulty is multiplied further in Angoff studies that estimate cut score for multiple student performance categories. Besides, estimating item difficulties requires conceptualizing items which arguably is also a pre-requisite for conceptualizing the target student population. Although conceptualizing test items is a pre-requisite for conceptualizing the target student population, however it has been given little attention. The current training practice presents isolated review of the knowledge and skills measured by the test and require participants to take the test as a mechanism to conceptualizing items. The practice of requiring participants to take the test arguably might foster instead processing of surface features of the test items as opposed to the constructs they measure.

Meanwhile, the training process for conceptualizing the target group of students often seeks to create a common conceptualization about the knowledge, skills, and abilities of the target group of students (Berk, 1986; Mills, Melican, & Ahluwalia, 1991; Reid, 1991). According to Giraud, Impara, and Plake (2005), this training process often employs an *a priori* definition of the target examinee which is often used for training discussion and directed by a facilitator. The discussion typically focuses on the knowledge and skills of the target group of students in relation to domain of the test for which a cut score is desired, and as defined by the performance standard. Fostering common understanding of the target group of students among participants which is the intended outcome of this training activity is hoped to be the prerequisite for agreement among participants on conception of target students, and thus to result in a cut score that is consistent across judges. Usually for the Angoff method training, following this group discussion aimed at target student performance category learning, there are instructions on how to conceptualize the target group of students.

Research efforts directed at reducing the cognitive complexity of the task of conceptualizing the target group of student by training instructions have introduced four types of instructions on how to conceptualize the target group of students, namely: (1) conceptualize a typical real student; (2) conceptualize a group of real students; (3) conceptualize a typical hypothetical student; and, (4) conceptualize a group of hypothetical students (Giraud, Impara, & Plake, 2005; Hein & Skaggs, 2010; Impara & Plake, 1997, 1998; Skaggs & Hein, 2011). Summarily, the training instructions for the task of conceptualizing the target group of students have asked participants to conceptualize either a single versus group of students; or, a hypothetical versus real group of students that barely makes it in the performance categories. Some of the more prominent of these research addressing cognitive complexity by instructions on how conceptualize the target group of students are reviewed in the immediately following paragraph.

In order to reduce cognitive complexity of the task of conceptualizing the target group of students, in their study, Impara and Plake (1997) proposed that the participants be directed to conceptualize a single real student who is known to them and who is typical of the target group of students. This approach based on conceptualizing a single real student aligns with the second component task of estimating the probability (or percent chance) that the student would correctly respond to the test items. The prescription was based on literature from studies that suggested that teachers are better able to estimate the performance of individual students in their class than the performance of the total group of students collectively (Hoge & Coladarci, 1989). The argument in favor of the prescription was that although an individual teacher may not be highly accurate in estimating the performance of a student on an item, the aggregated estimates across a collection of items and teachers might be quite accurate for the total group of students.

To support these claims, Impara and Plake reported a couple of empirical studies comparing the Yes/No Angoff method with the modified Angoff method. The studies were based on two variations of instruction on the conception of the target group of students. Specifically, the yes/no group was instructed to conceptualize a student in their current class who represented the target group while the modified Angoff group was instructed to conceptualize hypothetical 100 students from the target group. The finding from debriefing was that several of the teachers who used the modified Angoff method indicated they had quickly moved away from the difficult task of trying to imagine a group of hypothetical target students and had instead visualized one or two of their students. The conclusion based on the findings of this study was that the strategies used by the participants in conceptualizing the target group of students were similar and the only difference was in the use of probabilistic strategies for estimating the proportion correct with the modified Angoff method versus a more deterministic method for estimating if the target student would answer correctly or not with the Yes/No Angoff method.

Giraud, Impara, and Plake (2005) investigated how teachers who participate in cut score setting workshops conceptualize the target group of students, who barely make it in a performance category (e.g. proficient) in the Angoff standard setting method. They conducted two cut score studies, one based on mathematics test and the other based on reading. Their study asked the participants to think of a specific student whom they knew, that they would consider as barely making it in the performance category in estimating item difficulties. At the end of the study, the teachers were asked to provide a written description of the specific student who they had in mind that fits the performance description. Their finding was that teachers' descriptions of the target examinees for the both studies reflected the definitions provided by workshop facilitators. Also, the finding was that the descriptions of the mathematics group that used a more detailed performance level descriptor were more homogeneous. This study finding suggested that teacher participants might be able to adequately conceptualize the target group of students especially with a detailed performance level descriptor. The finding of this study leads to the speculation that the major source of challenge for participants of Angoff studies may be estimating the performance of this group of students on test items.

Hein and Skaggs (2010) research focused on the cognitive processes that occur during conceptualizing the target students and for their empirical study they directed participants to conceptualize an entire classroom of hypothetical group of students. In making a case for this prescription, they highlighted that conceptualizing an entire classroom group of hypothetical students was a simpler task than conceptualizing a single typical hypothetical person. Besides, their argument was that the aim of asking participants to conceptualize a hypothetical group of borderline students was for the participants to estimate the performance of an entire target population. Their supposition was that if participants were to reference only borderline students whom they had taught that it is possible that this narrower subset of the borderline students would not be representative of the target population. The empirical study compared the yes/no method and single passage book mark method[4]. For both studies, the participants were asked to conceptualize a classroom of hypothetical target group of students. Empirical evidence from data collected using in-depth focus group interview with eight participants from each of the panel meetings, and a whole-text analysis was that most participants experienced difficulties in attempting to conceive of an entire classroom of target students. Instead, most of the participants used the alternative cognitive strategy of thinking about only those particular students present in their classroom that fits the borderline performance description and the number of such students the participants thought of was as few as one and as many as six.

[4] A variation of the bookmark standard setting procedure for passage-based tests in which separate ordered item booklets are created for the items associated with each passage.

Skaggs and Hein (2011) research compared the passing scores resulting from the single passage variant of the bookmark method and the yes/no Angoff method. Their claim was that these two methods provided the most simplification of the bookmark and the Angoff method judgmental tasks, respectively for passage based tests. Therefore, their study was to compare judgments emerging from the methods and to test the cognitive complexity reduction hypothesis. In the empirical studies both groups of participants were asked to base their judgments on a hypothetical classroom of barely proficient students. Their rationale for defining the target population in terms of hypothetical classroom rather than the typical barely proficient student or 100 barely proficient students was that they considered the classroom unit to be more familiar to participants. Although this research did not directly test the hypothesis that the task of conceptualizing a target group of students in terms of a hypothetical classroom was simpler, however the point of reviewing this study was to highlight previous efforts at adjusting instructions on how to conceptualize the target group of students. Besides, this study provided the context for the prior reviewed study by Hein and Skaggs (2010), which directly explored the cognitive processes of participants of conceptualizing the target group of students.

The sparse empirical work investigating and describing the effects of training instructions and practice on the task of estimating item difficulties highlights the need for further research on the Angoff method training. The question then in need of researching is: what type of adjustment to the cognitive processing parameter of the probability judgment task through training instructions and practice activities would yield reliable and accurate probability judgments? To address this question however, it is necessary to understand the cognitive processes and strategies underlying probability judgment tasks a gap that this dissertation hopes to address. In the meantime, research on training feedback is reviewed next, followed in order by review of the alternative standard setting methods, previous research efforts addressing the cognitive processes and strategies of the Angoff tasks, the theoretical motivations for, and proposed approach of this dissertation.

3.2.2. Research on Training Feedback

Most of the attempts to address the cognitive complexity of the Angoff method tasks have emphasized feedback. As a consequence, most of what is known about the Angoff method training is about feedback. Hence, there currently exists different feedback information introduced to the Angoff method training in the effort to reduce the cognitive complexity of the tasks for participants, many of which are derivatives of empirical data namely, student responses on the test.

Reckase (2001) provided an organizational structure for the prevailing types of feedback. He conceptualized the types of feedback along a continuum that is straddled by two pivotal ends, presented in increasing order of cognitive effect as follows: (1) at the left end of the continuum is information provided to connect participant's ratings with observed student's performance, called normative feedback, an example of normative type of feedback information was the consequences feedback; (2) at the right end of the continuum is information that helps participants to decipher if they properly understand the standard setting process, called process-oriented type of feedback, examples of process type of feedback was the construct map feedback, a specific example of which is the Reckase's chart; and, (3) at the middle were multi-purpose forms of feedback that can either serve to connect participant's judgments with student performance or foster error correction, called the hybrid type of feedback.

The consequences feedback which is an example of the normative type of feedback shows the estimated percentage of examinees above a cut score or the distribution of scores on a test. The Reckase’s chart feedback which is an example of the process feedback, shows the possible test scores (aka latent ability score) and the relationship of these test scores to proportion of the target group of students that would respond correctly to test items when using item response theory (IRT) models. The Reckase’s chart feedback helps the participants to gain better understanding about how the students at each test score interacted with the test items so that they can decide if they want to change their item difficulty ratings to be consistent with those of a particular test score.

Many more examples of feedback used in Angoff studies fell into the hybrid category. Examples of the hybrid feedback were (1) the rater location feedback which shows the distribution of cut scores set by each participant with the each participant’s location indicated by a code letter only known to them; (2) the proportion of entire students or conditional proportion of students (e.g. above or below the cut score or at deciles), responding correctly to each test item; and, (3) group discussion of items (Reckase, 2001; Wyse, in press). The location of the hybrid types of feedback varies depending on whether the purpose of using them is to help the participants understand how the items function or if they serve as norm that serves to give them information about the performance of students.

Due to the greater emphasis on feedback in the Angoff method training, most of existing empirical evidence about training procedures directed at the task of estimating item difficulties is about the feedback procedure. Most of the existing evidence suggests that in the absence of performance data feedback, participants estimate of examinee performance are at best moderately correlated with item difficulties and suggestive of being flawed (American College Testing [ACT], 1995a, 1995b; Clauser, Swanson, & Harik, 2002; Clauser et al., 2009a; Cross, Impara, Frary, & Jaeger, 1984; Hanick, 1999; Impara & Plake, 1998; Norcini, Shea, & Kanya, 1988; Reckase, 2000). Perhaps, the latter evidence might have contributed to the great emphasis on performance data feedback. The immediately following paragraphs give a snapshot review of specific evidence about the effect of feedback on Angoff method training participants.

Reckase (2001) citing earlier research on the Angoff method feedback procedure highlighted mixed effect of feedback (Cross, Impara, Frary, & Jaeger, 1984; Norcini, Shea, & Kanya, 1988). Norcini, Shea, and Kanya (1988) found that providing item *p*-values (proportion correct) during the standard-setting process had no effect on the level of cut score. However, limitations of their study were that it used only six participants and a pre-post design that did not allow separation of effects of various parts of the process. Cross, Impara, Frary, and Jaeger (1984) compared Angoff method with two other standard setting methods and found a significant effect of item *p*-value information on ratings. The explanation Reckase offered for the mixed results on feedback was that these research studies employed different contents of tests, sample sizes of participants, types of feedback, and levels of control.

There are however, some consistent findings about effect of feedback in the Angoff standard setting literature. Reckase (2001) citing work done by ACT to guide the design of standard-setting process for NAEP (e.g. ACT, 1995a, 1995b) reported a couple of consistent findings about effect of feedback: (1) process feedback has the typical effect of reducing the standard deviation of cut scores set by participants; and, (2) the effect of feedback reduces as the number of feedback rounds increases. Cizek (2001) also reported the following consistent findings (1) discussion and impact data results in consistency of cut scores; (2) there is the tendency for participants item ratings to converge towards the item *p*-values presented to them regardless of whether the estimates were based on the total group of students or a stratum selected to approximate the hypothetical student group.

Wyse (in press) citing a number of research studies on feedback reported that use of item performance data can help participants to reduce their inconsistency, but that it does not necessarily completely remove inconsistency (Brandon, 2004; Busch & Jaeger, 1990; Clauser et al., 2002; Clauser et al., 2009a; Clauser, et al., 2009b; Clauser, Mee, & Margolis, 2011; Margolis, 2011; Mee, Clauser, & Margolis, 2011). For instance, Busch and Jaeger (1990) found a group correlation of 0.60 of item ratings with the overall item scores without performance data and a correlation of 0.84 after receiving performance data; Clauser et al. (2009a) found group correlations of around 0.60 of item ratings with conditional expected item scores without performance data and correlations in the 0.90's after providing performance data feedback. These findings are also consistent with those reported by Brandon's (2004) in his review of research on the Angoff method. Recently, Wyse (in press) offered an alternative explanation of restriction of range phenomena for the robust finding in the literature of the feedback effect of increasing correlations and decreasing standard deviations of judgment outcomes across feedback rounds. Therefore, Wyse suggested the need for research to reconsider the use of the correlation index for evaluation.

There is also evidence for preference for the process-oriented type of feedback. Reckase (2001) reported work done by ACT that suggested that participants preferred Reckase's chart feedback (Hanick, 1999).Wyse (in press) citing several studies by ACT reported findings based on survey questions that suggested the Reckase's chart was a useful feedback mechanism in NAEP (ACT, 2005; Hanick, 1998a, 1998b, 1999; Loomis, 2000; Yang, 2000). Wyse (in press) provides a case study of a National Assessment of Educational Progress (NAEP) Angoff standard-setting process that used Reckase charts (Reckase, 2001), and his results suggested that the ratings for the second and third round, where Reckase charts was introduced showed improvement in cut scores. Besides, the findings showed increased correlation of the ratings with the conditional expected item scores at the estimated cut-scores and decreased discrepancies between item ratings and estimated cut scores.

Some interesting lines of inquiry about the impact of feedback that have the prospect of enhancing understanding of the Angoff method and that deserve studying further were also identified in the literature. For instance, a research project studied the extent to which item performance data feedback was process-oriented by introducing erroneous data to the training. The findings suggested mechanical reliance on these data even when they are erroneous (Clauser et al., 2009b). Wyse (in press) reported a new line of inquiry on the Angoff method on the impact of mode of presentation of feedback data and the instructions on how to use them on cut-scores. The findings from this line of research suggest that instructions have the following impact: (1) participants are less likely to use feedback data when instructed that they are faulty; (2) the format (e.g. frequency versus proportion correct) in which feedback data is presented did not necessarily impact cut-scores; and, (3) more detailed feedback information resulted in greater correlations of the panelists' item ratings with conditional expected item scores than less detailed information (Mee, Clauser, & Margolis, 2011).

Reckase (2001) highlighted the shortfall of the reviewed prevailing Angoff method training feedback as having to do with lack of instruction on their meaning and on strategies to integrate them into judgments. He highlighted that Angoff standard-setting researchers often acknowledge that participants need training to conceptualize a person who barely makes it in a performance category and to estimate item difficulty for the group. However, researchers pay little attention to the need to train participants about the meaning and use of feedback even though they are unfamiliar with the feedback they receive. It is only recently that some research studies have begun to investigate how the format for providing these feedback data and the instructions on how to use the information impacts cut scores (Mee, Clauser, & Margolis, 2011).

However, with the introduction of feedback data training implies three types of training for the Angoff method in serial order: (1) training to conceptualize the target group of students; (2) training to estimate item difficulties prior to feedback; and, (2) training to integrate feedback data in revising item difficulties. Although introduction of instructions on the use of feedback data is apt however, it would mean less efficient training. Moreover, training participants on how to change their recommendations to reflect quantitative feedback information might mean introducing instructions based on more complicated probability principles such as the Bayes rule to the Angoff training process. Also, as Reckase highlighted the introduction of multiple rounds of feedback would require that decisions be taken about the ordering of the feedback information because different ordering may result in qualitatively different effects on the participants.

In all, the identified shortfall of the prevailing feedback approach to addressing cognitive complexity of the Angoff method tasks is that they serve to present the participants with a possible representation of the final outcome of judgment, are presented without reference to the knowledge and skills requirements of the tasks, and on how to use them to revise judgments. Use of a lot more information demands more complex cognitive strategies for integrating the information into judgment (Tversky & Kahneman, 1971, 1973, 1974). However, these feedbacks are often provided without instructions to improve conceptions and on how to use them to revise judgments. Because these prevailing feedback approaches serve to present the participants with possible representation of the final answer of the judgment while instructions are not provided that references the knowledge and skills requirements of the tasks and to make use of the data process oriented they at best hybrid in nature.

Consequently, this dissertation proposed a different categorization scheme for feedback: final-outcome or intermediate-outcome. The final-outcome feedback are all quantitative information that provide a possible representation of the outcome of the task of generating proportion of correct for the target group of students. The intermediate-outcome feedback addresses the pre-requisite knowledge and skills requirements of the Angoff method tasks. Specifically, the intermediate outcome feedback serves to enhance conceptual understanding of the participants of the knowledge and skills constructs measured by the test items, the target student population, and instructs the participants about process to integrate the information in generating the proportion correct for the target students. The feedback types that are admissible in this criterion-referenced to knowledge and skills requirements of the Angoff method tasks feedback training are discussions of the intermediate outcomes and the empirical data types augmented with instruction that reference knowledge and skills requirements to make them process-oriented.

3.3. Research on Alternative Standard Setting Methods

The flip side approach to multiple feedback procedures to addressing cognitive complexity of task of estimating item difficulties is the introduction of alternative standard setting methods. Numerous standard setting methods have being proposed with purported simplification of the Angoff task (Raymond & Reid, 2001).The motivation behind most of the methods was to reduce the cognitive complexity of this task either by reducing or eliminating the need for participants to generate probabilities of correct response for the borderline subgroup of examinees (Hein & Skaggs, 2010). An example of method introduced in response to cognitive complexity of this task was the bookmark method (Mitzel, Lewis, Patz, & Green, 2001). The Bookmark method is currently the method of choice in the public school settings (Karontonis &

Sireci, 2006). As a consequence, in order to simplify this review, the Bookmark method would be used for illustration of the limitation of the alternative method approach to addressing the complexity of the task of generating item difficulties.

In the Bookmark method, items are ordered in terms of their item response model estimated difficulties and the task requires participants instead to make only one decision about the entire test. The decision is to place a bookmark on an item that separates the test items into two groups, those the minimally acceptable examinees can answer correctly with at least a specified probability, and those they can answer correctly with less than the specified probability. The probability used for ordering the test items in terms of item response model estimated difficulty is called the response probability (RP). Subsequently, the item response model estimated difficulties of the items located at the bookmarks of participants are averaged to estimate the cut score.

Although the Bookmark method may reduce cognitive complexity, it is arguable that it does so by trading accuracy for simplicity (Hein & Skaggs, 2010). For instance, some researchers have pointed out that the Bookmark method suffers from response probability (RP) indeterminacy (Haertel & Loriè, 2004). Cizek and Bunch (2007) have also observed that in order for the Bookmark procedure to be accurate there should be a large number of items in the ordered item booklet(OIB) that are near the location where the participant intends to set their cut score. Research studies have also demonstrated that the bookmark method can result in higher potential cut score biases than the Angoff method. The prominent of these research studies comparing the technical properties of the Bookmark and the Angoff method are reviewed as follows.

Reckase (2006) study of the psychometric theory for standard setting using simulations based on the Rasch and 3 PL models, investigated the potential statistical bias in a single panelist's intended cut score with the Bookmark procedures. Results showed the potential impact that gaps in the difficulty between items could have in the Bookmark procedure. This study showed that participants' cut score was recovered more accurately with the Angoff method than the Bookmark procedure. The study also suggested that depending on the location of the participant's desired cut score, the bookmark method could result in a large amount of statistical bias (Reckase, 2006). Wyse (2009) study applied comprehensive item response evaluation indices based on residuals and absolute residuals at the participants intended cut score to determine the potential cut score biases produced by the Bookmark and the Angoff method. The finding by Wyse (2009) showed that the bookmark method has the potential of leading to biased cut scores due to possibility for gaps in the score scale from lack of standard setting stimuli at every score scale location.

Besides these highlighted bits of research evidence, the argument of this dissertation is that by relying on item response models, the bookmark method is less ecologically valid because there are potentially many models that could be fit to item response data and each with potentially different ordering of items in terms of difficulty. Moreover, the argument against the alternate standard setting method approach to addressing cognitive complexity of the task of estimating item difficulties, is that by tweaking response sets, these methods although may result in outcomes that are more consistent, however may be less veridical. Therefore, the approach to addressing cognitive complexity of the task of estimating item difficulties in this dissertation is through CTA to understand the knowledge and skill requirements of the Angoff tasks and application to the design of training. The next section reviews previous research on the cognitive processes underlying the Angoff method tasks.

3.4. Research on Cognitive Processes Underlying the Angoff Method Tasks

Research on training in the cognitive psychology literature suggests that optimal design of training demands a needs analysis to understand the mental processing requirements of task performance (Campbell, 1971; Goldstein, 1980; Latham, 1988; Salas & Cannon-Bowers, 2001; Tennenbaum & Yukl, 1992; Wexley, 1984;). Raymond & Reid (2001) conducted the foundational research on cognitive task analysis for the Angoff method, however their work highlighted the component tasks but fell short of uncovering the fundamental cognitive processes and strategies for performing the task of estimating item difficulties.

Since then, McGinty (2005), a research work on which this dissertation builds on, called for the need to uncover the "Black box" of standard setting methods, the factors and processes that participants consider in their judgment. Recently a new wave of research initiatives apparently responding to the need for research to understand cognitive processes of standard setting methods has begun. This line of research relaxes the assumption that participants understand the task they are asked to perform and are able to carry out the procedure accurately (Wyse, 2009). They seek instead to understand the mental processes underlying the standard setting methods especially the Angoff method by conducting research to illuminate the thought processes and experiences of participants during standard setting (e.g., Buckendahl, 2005; Ferdous & Plake,2005; Giraud & Impara, 2005; Giraud, Impara, & Plake, 2005; Hein & Skaggs, 2010; Skaggs & Hein, 2011).

However, most of this research explored the cognitive processes underlying the first component task of the Angoff method, the conception of borderline group of students. The more relevant of this work was reviewed in the previous section on training, in order to avoid redundancy they would not be reviewed here again. However, it is important to recap that most of these studies uncovered that irrespective of instruction that participants engaged in more fundamental cognitive processes of simply relating the performance descriptors to actual students in their classrooms instead of trying to imagine some abstracted notion of hypothetical group of students. This finding lends support to exemplar approach to categorization being in play in Angoff standard setting instead of the prototypical approach to categorization. Although these latter research initiatives are beginning to illuminate some cognitive strategies but these efforts are still at the descriptive level. Besides, none of these studies revealed the cognitive strategies involved in the task of estimating item difficulties, however based on the findings about the first component task one can infer that more fundamental strategies might also be in play.

The argument in this dissertation is that standard setting research can do even much better than mere description of cognitive processes involved in the Angoff tasks. Therefore, this dissertation assumes the rightful role of educational research field of prescription. The interested reader can refer to Chapter two for types of cognitive process theories and the roles of different fields of research. This dissertation draws mostly from the cognitive psychology descriptive theories of the cognitive processes underlying the task of judging conditional probabilities in prescribing heuristic process for this aspect of the Angoff task. The section that immediately follows delineates the motivation for the proposed new direction for Angoff method research targeted at addressing the cognitive complexity of its task by training while maintaining the goals of veridicality of outcomes and efficiency of implementation of laboratory Angoff studies.

3.5. Motivation for the Proposed Angoff Method Heuristic Training Paradigm

This section briefly reviews the theoretical basis of the modified Angoff heuristic training paradigm proposed by this dissertation.

This dissertation proposes a comprehensive approach to addressing the cognitive complexity of the Angoff task that attends to its two component tasks. Specifically, a heuristic training paradigm is proposed for the Angoff method. The assumption underlying the proposed heuristic training paradigm is that human participants of Angoff studies have limited capacity to process different types of information in their judgment. This dissertation argues for human processing limitations as the factor delimiting utility of the Angoff method. Therefore, the rationale for the proposed heuristic training paradigm is to simplify the cognitive processing parameters of the Angoff method tasks, which in serial order are: (1) conceptualizing the target student group that barely makes it in a performance category; and, (2) judging conditional probability of correct response for test items. The heuristic paradigm explicates the Angoff tasks by breaking them down into simpler mental operations.

From an educational measurement perspective, the proposed heuristic principles were based on criterion-referencing so that categorical domain construct predictors of item difficulty are given maximum weight in the processing instructions while construct irrelevant factors are given little or no weight. Therefore, the proposed heuristic training is based on constrained categorical test construct information processing. Put differently, the heuristic training does not constrain the sample size of the students, items, and testing episodes to be considered in the judgments, rather it constrains the type and number of features of the items and students to be considered. The types of features are the categorical content domain features of students and items and the number is constrained by the constructs measured by the test items. The proposed heuristic training was predicated on the assumption that test items used for Angoff standard setting are selected that are adequate measures of test constructs and that have minimal construct irrelevant properties that can influence item difficulty. The motivation for proposing the Angoff training, based on heuristic principles is to: (1) reduce cognitive complexity of the tasks; and, (2) increase veridicality of probability judgments of laboratory Angoff studies; and, (3) increase efficiency in practical implementation of laboratory Angoff standard setting studies.

The proposed prescription for the first component task directed participants to think about actual students in their classroom. Citing Impara and Plake (1997), this cognitive process strategy was informed by literature from studies that have examined the accuracy with which teachers can estimate the performance of individual students in their class rather than the performance of the total group of students collectively. The argument for this strategy was that although an individual teacher may not be highly accurate in estimating the proportion of their students who fit the performance description that would answer each item correctly, however, the aggregated estimates across a collection of items and teachers might be quite accurate (Cooper, 1995)**.** Much of this literature has been summarized by Hoge and Coladarci (1989). The reviewed standard setting research pertaining to different instructions about how to conceptualize the target group of students suggested that teachers are better able to think in terms of real rather than hypothetical group of students.

Hence, the latter evidence suggested that teachers stored information about student performance categories may be in terms of specific examples of students instead of some abstracted representation of students called prototype. Meanwhile, the cognitive psychology literature on categorization suggests that the exemplar approach is more effective than the prototype approach because it contains more information about a category. Because of the latter evidence and the reviewed standard setting research evidence that suggested that teacher participants find it easier to think in terms of real examples of students than an abstracted representation, this dissertation prescribed thinking in terms of real examples of students. Also, the reviewed standard setting research pertaining to instructions about how to conceptualize the target group of students suggested that teachers are better able to think in terms of specific students in their classroom than about a group of students. Although the evidence suggests that teachers are limited in terms of thinking about a group of students, the prescribed approach in this dissertation was to have participants to conceptualize a group. However, the group size was left open for the proposed training for the reason that it might vary for participants because they may have experienced students of different ability population composition. Moreover, it is conceivable that the group of students' formulation might be a more simplifying approximation than a single student because it might involve recalling only one testing episode to estimate the relative frequencies. In addition, the rationale for having teachers conceptualize a group was because of the goal of accuracy and also because this conception maps well with the second subtask operationalized in terms of estimating the proportion of the identified group of students that would respond correctly to each test item.

The proposed prescription for addressing cognitive complexity of the second component task retains the task as judgment of conditional probabilities of correct response to test items by the conceptualized target students while prescribing short cut strategies called heuristics for performing the task. The prescribed strategies for judging the proportion of the target students that would respond correctly to each test item, once the target group have being conceptualized were informed by cognitive psychology literature.

Research in cognitive psychology suggests that because of limits on computing speed and power, and task environment constraints that people rely on approximate methods called heuristics that reduce the task of judging probabilities into simpler mental operations. According to Todd and Gigerenzer (2007), these heuristics are quite economical and effective, and can yield the right answers when applied in appropriate environments. Because different heuristic approximations give rise to different solutions to the probability judgment problem, it is important to construct a theory of its processes and to describe the environments to which it is suitable (Simon, 1990).

Given the striking analogy between the research field in cognitive psychology on probability judgment and the Angoff method tasks, it is deemed appropriate to draw from the ample cognitive processes already uncovered in this research area and forge ahead with the task of prescription. The premise is that the heuristic principles are adequately adapted to the public school Angoff standard setting environments. Consequently, this dissertation research addresses the cognitive complexity of the task of estimating item difficulties by reformulating the task in terms of the more fundamental heuristic cognitive processes. The heuristic principles were operationalized by the proposed heuristic training program. Adaptability of the heuristic principles to the public school Angoff standard setting environments was predicated upon the assumption that potential participants of Angoff studies in this context, the classroom teachers are knowledgeable about empirical facts pertaining to the interaction of students with test items. To the extent that this premise holds true, then the proposed training paradigm besides increasing efficiency, has the potential of creating the required balance between cognitive complexity, veridicality, reliability, and accuracy of Angoff method probability judgment outcomes.

For instance, the heuristic training paradigm accomplishes reduction of cognitive complexity through instruction on restricted categorical content domain construct information processing; increases veridicality by instructions to consider real instead of hypothetical experienced instances of student and item categories and testing episodes; increases reliability through homogenizing heuristic strategy instructions that could potentially reduce within item category and between participants judgment variance; and increases accuracy through instructions, that draws on the tenets of the central limit theorem, on the most predictive judgment cues, and on unrestricted explicit memory search for experienced students, items, and testing episodes. Besides, the proposed heuristic principles are quite efficient and economical (Tversky & Kahneman, 1974).

The cognitive psychology probability judgment heuristic research paradigm suggested two heuristics namely: representativeness and availability heuristics that mediates most probability judgments (Tversky & Kahneman, 1974). The representative heuristic mediates probability judgments through similarity judgments while the availability heuristic mediates probability judgment by the recall of what comes to mind first. The interested reader can refer to Chapter two for details of the probability judgment heuristic paradigm. The fundamental cognitive tasks underlying the representativeness and the availability heuristics are categorization and recall, respectively. Although these are prescribed for the task of estimating item difficulties, however these two sub-tasks also apply to the first task of conceptualizing the target group of students. Meanwhile, the heuristic mechanism for this dissertation is mainly categorization through similarity judgments. The instruction for categorization constrained the features of categories to consider. However, the recall task was not simplified so that the expectation is that participants engage in extended memory search in generating experienced category members and response data. The rationale for not simplifying the recall task by asking the participants to recall a single student or item or test taking episode is because it is considered that the more members of a category and episodes that can be recalled the higher the chances for accuracy of outcomes.

To study these hypotheses this dissertation employed participants in two training studies that recommended cut score for the Proficient student performance category. The training based on the heuristic principles was called the Heuristic training. The training based on the typical modified Angoff method instructions was called the Normative training.

For the Normative training, the prescription for the first component task was to think of a hypothetical group of the target students that barely makes it in the Proficient performance category, while the prescription for the second task simply asked them to estimate the proportion of the students that would respond correctly to each test item.

For the Heuristic training, the prescription for the first component task was to think of actual students in their classrooms that fit the description of barely proficient, while the prescription for the second task asked them in serial order to: (1) categorize each test item based on their content domain knowledge and skills features (e.g. content strands and depth of knowledge levels) and similarity judgments; (2) think of items they had experienced that fit that category; and, (3) estimate the proportion of their students that fit the barely proficient description that were able to respond to those items.

The Heuristic and the Normative group instruction for the first component Angoff task differ in terms of real versus hypothetical student, respectively while both instructions prescribed thinking about group of the barely proficient students. The second component task instruction for the two groups differs qualitatively because it was explicated for the Heuristic training to indicate the intermediate steps of categorization and recall operations to generating proportion correct for the target barely proficient students. However, the Heuristic training instruction for the second component task was left open with regards to the number of similar experienced items the participants are to recall and how to combine proportion of correct estimates if more than one item is recalled. The rationale for the latter is that it was perceived that there is a limit at which task explication might become counterproductive and therefore seize to simplify the process especially when participants have different experiences.

To conclude this section, just as the distinguished social psychologist Gordon Allport "memorably noted, the human mind must think with the aid of categories. We cannot possibly avoid this process. Orderly living depends upon it" (Fryer & Jackson, 2007, p. 3). Likewise, the Heuristic training embodies the principle of categorization for the Angoff standard setting task and holds promise of resolving the cognitive complexity limitation of the task while increasing efficiency, accuracy, veridicality, reliability, and accuracy of judgments.

Chapter Four: Re-conceptualization of the Angoff Method

This chapter integrates ideas from the reviewed theoretical frameworks and the Angoff standard setting method literature in explicating on the Heuristic training method to addressing cognitive complexity of the task of judging conditional probabilities of correct response. There are four sections in this chapter. The first section elaborates on potential cognitive and non-cognitive factors that could influence the probability judgment outcome in accordance with the Heuristic training paradigm[5]. The second section delineates the knowledge and skill requirements of the modified Angoff heuristic tasks. The third section presents the heuristic cognitive process model for the Angoff task. The reader should please note that the model is for the second component task of judging conditional probabilities of correct response to test items when the target group of students has being conceptualized. However, it is important to point out that the explicated cognitive processes of the model could equally apply to the first component task of conceptualizing the target group of students. The fourth and the last section expound the assumptions of the heuristic cognitive process model.

4.1. What Factors Influence Participants Probability Judgments?

First and foremost, it is important to recap the theoretical underpinnings of the ensuing conceptualizations: (1) The realist view of measurement and the frequentist probability schools of thought were adopted in this dissertation; (2) The view of standard setting adopted in this dissertation is the parameter estimation model an example of which is the Reckase’s (2009) model of standard setting. According to Reckase’s (2009) model, standard setting is more appropriately called standard translation. He used standard translation in a metaphorical sense to

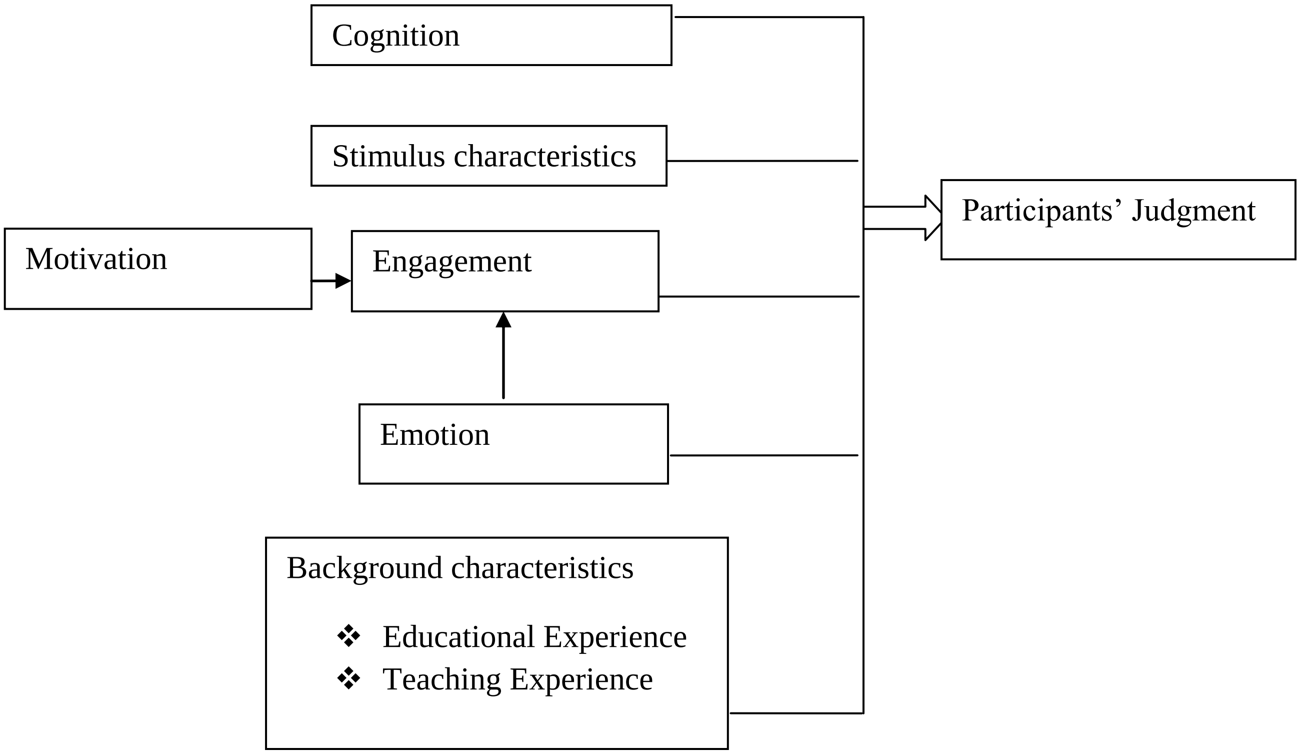
[5] Probability judgments is also known as relative frequency estimates or classical item difficulty or p-values

draw analogy between the standard setting process of translating the language of policy to a numerical test score and that of translating text from one language to another; and, (3) the bounded view of human rationality and the heuristic probability judgment paradigm were adopted to the study of the Angoff method. The interested reader can refer to Simon (1957) for the bounded notion of human rationality. The aforementioned theoretical views are exemplified by the heuristic cognitive process model for the Angoff tasks and the heuristic training paradigm, directed at reducing cognitive complexity of the translation process.

Consequently, it is here hypothesized that participants' performance of the Angoff method standard translation task, requiring probability judgments is a function of their background characteristics, cognition, stimuli characteristics, and non-cognitive personality attributes. The precise non-cognitive personality attributes considered were: motivation, emotion, and engagement. It is appropriate at this juncture to provide the working definition for the non-cognitive personality constructs. These definitions are as follows: Motivation refers to reasons underlying behavior (Appleton, Christenson, Kim, & Reschly, 2006). Ryan and Deci (2000) distinguished two main types of motivation namely: intrinsic and extrinsic motivation. Intrinsic motivation is defined as doing an activity for its inherent satisfaction rather than for some separable consequence, while extrinsic motivation pertains to doing an activity to attain some separable outcome. Emotion relates to participants affective reactions including interest, boredom, happiness, and anxiety about the tasks or while performing the task. Engagement refers more to behavior and reflects a person's active involvement in a task or activity (Appleton et al., 2006). Motivation and positive emotion are necessary but not sufficient for engagement. For instance, one can be highly motivated and in a good emotional state but not actively engage in a task. The constructs motivation, engagement, and emotion were considered because there is documented evidence in the cognitive psychology judgment and training literature that they are predictive of optimal task performance (Campbell, 1971; Dion, Berscheid, & Walster, 1972; Goldstein, 1980; Johnson & Tversky, 1983; Latham, 1988; Lichtenstein, Slovic, Fischhoff , & Combs, 1978; Loewenstein, Weber, Hsee, & Welch, 2001; Salas & Cannon-Bowers, 2001; Schwarz & Clore, 1983; Tennenbaum & Yukl, 1992; Wexley, 1984).

The hypothesized relationship between these factors with probability judgment outcomes is shown in Figure 4-1.

Figure 4-1: Conceptualized Factors Influencing Performance of Participants



The arrows pointing right in Figure 4-1 indicate the order of dependencies of the variables. It is proposed that background characteristics be addressed through selection; stimuli factors by selection and design; cognition through training; motivation by incentives and social accountability; and, engagement and emotion although may not explicitly addressed by any component of the process be measured and taken into consideration when interpreting the results of the study. Meanwhile, the relatively novel notion of social accountability emanates from the theoretical assertion that motivation is enhanced by social benefits such as need to make a favorable impression and avoid embarrassment (Larrick, 2004). For instance, requiring participants to explain the rationale behind their recommendations to others during discussion can foster social accountability and therefore enhance motivation. In the latter case, the principal mechanism by which social accountability improves judgment is through preemptive self-criticism. The notion of preemptive self-criticism suggests that in preparing to justify judgments to others, participants would anticipate flaws in their arguments thereby improving their judgment processes and outcomes.

4.2. Modified Angoff Task: Knowledge and Skills Requirements of Heuristics Strategy

A Modified Angoff method in the standard setting literature is currently conceived of as training procedural modification to the Angoff method (Cizek, 2001; Cizek & Bunch, 2007; Zieky, 2001). Usually the procedural modification is introduction of a different feedback to training and without due reference to the knowledge and skills requirements of the tasks. However, this dissertation re-conceptualizes modified Angoff methods as adjustments to cognitive factors and processing requirements of the Angoff method tasks. These adjustments are more appropriately operationalized through training instructions on what factors to consider in judgments and on how to integrate them in performing the Angoff tasks.

Hence, training as conceived in this dissertation addresses both the knowledge requirements of the Angoff tasks and strategy for integrating the knowledge into judgments. The emphasis of the heuristic strategy is on the knowledge, skills, and ability categories of items and persons and their interaction in producing relative frequencies of correct response. The rest of the discussion in this section presents knowledge requirements, while the next section presents heuristic model for the Angoff task. Table 4-1 presents the results of cognitive task analysis (CTA) based on reviewed cognitive psychology research literature on probability judgment. The CTA result presented in Table 4-1 were also informed by Raymond and Reid’s (2001) foundational work on the CTA of the Angoff tasks. Before discussing the information presented in Table 4-1, it is important to highlight that the CTA results generalize directly to public school accountability standard setting contexts.

In Table 4-1, the first column shows the major knowledge and skills requirements of the probability judgment heuristic, the second column gives the pre-requisite knowledge and skills required for participation in a modified Angoff heuristic training study, and the third column presents training activities that augment knowledge and skill deficiencies of participants. By far, the most important pre-requisite knowledge and skills are empirical knowledge of student population and their interaction with tasks in the content domain. It is necessary that at least some of the participants are knowledgeable about the student population and their interaction with test items, because discussion and elaboration of performance level description cannot completely augment for total ignorance of empirical facts. On the other hand, extensive knowledge of probability theories and axioms is a plus but not essential based on the heuristic perspective.

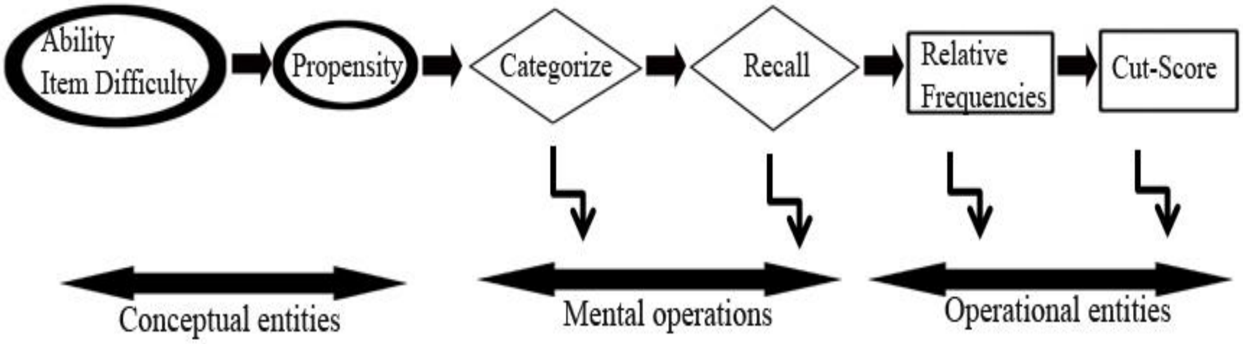
Table 4-1: Results of Cognitive Task Analysis of the Modified Angoff Method

Knowledge and skill requirements	**Selection factors**	**Training activities**
1. Substantive knowledge-knowledge of the content domain; purpose of test, test specification knowledge and skills constructs; Item content characteristics that influence difficulty (e.g. content strands, GLCE, and depth of knowledge categories)	Knowledge of content domain of the test; experience using or administering items of the content domain of the test	Discuss purpose and rationale for standard setting; explain background of test, test development, and item writing procedures; discuss construct relevant factors that influence item difficulty
2. Empirical knowledge - Examinee population, knowledge and skill attributes of students at each performance category	Taught or currently teach student population; Knowledge of levels of proficiency in examinee population	Elaboration of performance level description
3. Empirical knowledge-Interaction of students with exemplar items measuring the knowledge and skills constructs of the content domain; test performance data	Taught or currently teach student population; Knowledge of levels of proficiency in examinee population ; Declarative knowledge of interaction of students with tasks in the content domain; Knowledge of student cognitive ability and test performance; ability to recall student performance data	Elaboration of performance level description
4. Probability knowledge - Basic understanding of relative frequency scale and how estimate relative frequencies or proportions	Elementary knowledge of probability such as that proportion is a number between 0-1 and how to estimate proportion	Heuristic instruction and practice estimating item difficulty with feedback and discussion

4.3. Modified Angoff Task: Heuristic Cognitive Process Model

Figure 4-2 presents the heuristic cognitive process model for the Angoff task of judging conditional probabilities of test items when the target students have been conceptualized. The model is considered to be capable of representing the property of interest, the probabilities of correct response to test questions for the subpopulation of students at the threshold level (s) of performance in terms of relative frequencies of correct response to the test questions (Nichols, Twing, Mueller, & O'Malley, 2010). The model builds on McGinty's (2005) information processing model for the standard setting tasks. McGinty's model for standard settings was conceived in terms of a computer metaphor, to include inputs, processes, outputs, and consequences. In relating the information processing model to the task of translating performance level descriptions, operations of participants are distinguished from those of researchers. Specifically, operations of participants are unobservable cognitive processes that mediate the relationship between Angoff method stimulus inputs (that include performance level descriptions and test questions) and the outputs, and that direct operations of researchers. Thus, the modified Angoff method task is expressed in Figure 4-2 as a heuristic process model to include fundamental cognitive operations of participants, aimed at reducing cognitive complexity as follows:

Figure 4-2: Model for the Task of Translation of Performance Standards Using the Modified Angoff Method.



In Figure 4-2, conceptual entities are in ovals, cognitive processes are in diamonds, and operational entities are in boxes, with connecting arrows pointing right showing the direction of dependency of the postulated relationships. Also, downward pointing arrows are attached to the cognitive processes and operational quantities. The latter arrows indicate that the cognitive processes can encounter hindrance or be influenced by constructive processes, hence leading to errors in the operational quantities. The model postulates that the participants take as input the constructs that are the basis of measurement that is conceptions of threshold student ability groups, item difficulties, and associated propensities, with the latter, being the conceptual version of probability based on the frequentist account. Subsequently, by engaging in interactive cognitive processes of categorization and recall, for each test item, they generate operational versions of probabilities, the relative frequencies. These relative frequencies are translated to the operational version of the representative ability that is, the cut scores. Precisely, the cut scores are typically determined based on generalized binomial modeling principles as the sum of the test items estimated relative frequencies of correct response for the threshold student groups.

Because the focus of this dissertation is on addressing the cognitive complexity of the Angoff task, it is therefore appropriate to explicate further the cognitive parameters of the heuristic model.

- First, the model posits that test questions for which event of correct response for students who barely make it in a performance category are to be predicted are first classified with set of similar test questions and that the probability estimate is based on prior experience and explicit past relative frequency count of interaction of these students with related test questions[6].

- Second, the notion of categorization as expressed in the model is based on item feature matching similarity judgments. Feature matching theory is formulated in terms of the set theoretical notion of matching function rather than in terms of the geometric concept of distance. Therefore it is neither dimensional nor metric in nature (Tversky, 1977). The term feature as appears in the label of the theory denotes the value of a binary variable or a nominal variable. The feature matching approach to categorization assumes that test items are mentally represented as a collection of binary or nominal features so that similarity is described as a feature matching process. Therefore, categorization is constrained with respect to specification of binary or nominal features of test items to be considered in the similarity judgments. In its instantiation for the Heuristic training, the specified features were the constructs measured by the test items (e.g. Depth of Knowledge levels, content strands, and Grade Level Content Expectations). It is assumed idealistic heuristic principle, that the feature matching process gives constructs measured by the test items maximal weight, while construct irrelevant features of the items are given minimal weight in the judgment of similarity.
- Third, recall is not necessarily constrained by the model so that to the extent that extensive search of declarative and empirical information about test items and their interactions with the students in generating responses then the better the accuracy of judgments generated by the model.

[6] The reader can refer to (Keren, 1991, pp. 229-230) for support of plausibility of this cognitive strategy in the case of related events and based on frequentist interpretation

- Lastly, to put it differently, the Heuristic training does not constrain the sample size of the students, items, and testing episodes to be considered in the judgments, rather it constrains the type and number of features of the items and persons to considered. The types of features are the categorical content domain features of students and items and the number is constrained by the constructs measured by the test items.

4.4. Model Assumptions

As Raymond and Reid (2001) recommended, it is necessary to identify and test assumptions underlying any standard setting method including about the materials, participants, and activities. In response to this recommendation, this section presents the assumptions underlying the heuristic model for the Angoff task and organized in terms of materials, participants, and activities. However, the reader should please note that most of the assumptions about the participants were relaxed and addressed by the Heuristic training.

4.4.1. Materials

The assumptions in this section are about the stimuli used for the modified Angoff standard setting.

- The PLDs adequately describes the knowledge, skills, and abilities of the student performance categories intended by the policy makers
- The ability distribution of the performance category is unimodal
- The test measures a unidimensional knowledge and skills construct
- The test items are selected that are adequate measures of the content domain knowledge and skill constructs, with minimal construct irrelevant features that can impact item difficulty
- The items on the test represent the knowledge, skills, and ability constructs of the PLD

4.4.2. Participants

The assumptions in this section are about the knowledge, skills, and abilities of participants. Also presented are assumptions about non-cognitive personality attributes of the participants namely: about motivation, engagement, and emotional states of participants. It is assumed that participant:

- Have sound working and declarative memory
- Have teaching experience
- Are knowledgeable about the substantive content domain
- Are familiar with abilities of the student population
- Are knowledgeable about empirical facts pertaining to the interaction of the students with items that measure the knowledge and skills of the PLD
- Possess basic knowledge that proportion or relative frequency is a number between 0 and 1
- Can apply the heuristic strategy
- Are intrinsically motivated, adequately engaged, and emotionally stable

4.4.3. Activities

The assumptions in this section are about mental activities and training procedures. Mental activities are also included as assumptions because even when explicit instructions are given to participants, it is not guaranteed that they would follow them.

A. Mental Activities

It is assumed that:

- Modified Angoff judgments depend on empirical facts
- Modified Angoff judgments are a function of categorization and recall.
- Categorization is a function of item feature matching similarity judgment

- The feature matching process is neither metric nor dimensional
- The feature matching process is based on binary and ordinal level knowledge and skills constructs features of items
- The construct irrelevant features of persons and items are given minimal weight in the feature matching process

B. Training Procedures

It is assumed that:

- The heuristic training instruction is capable of yielding relative frequency estimates that correspond to the relative frequencies of students responses based on test calibration
- The performance standard elaboration process can augment deficiencies in knowledge of empirical facts about students and items
- Training practice on categorization constitutes deeper conceptual processing than taking the test and can enhance better recall and probability judgment performance
- Intermediate-outcomes feedback types can enhance conceptual understanding and yield more valid judgments than final-outcome feedback types of information
- Iterative feedback rounds of judgments may lead to normative influence such as technical adjustments of recommendations and to phenomenon called group discussion induced polarization effect[7].
- The Heuristic training on limited test construct information processing is capable of reducing cognitive complexity of the Angoff task of judging conditional probabilities

[7] Technical adjustment to recommendations implies either increasing or decreasing recommendations by a constant amount and in relation to norms

- The Heuristic training on test construct information processing can yield more reliable and accurate judgments than the Normative training based on individuating feedback information
- The Heuristic training is more efficient than the Normative training and can result in reduced mental effort and cost of execution of modified Angoff standard setting process

Chapter Five: Analytic Framework

5.1. Training Evaluation Criteria

In this dissertation, training is to be evaluated in accordance with Kirkpatrick's (1994) comprehensive framework for evaluating criterion referenced corporate training. Kirkpatrick's framework was identified as the most popular framework for evaluating training programs in the cognitive psychology training literature (Salas & Cannon-Bowers, 2001). Kirkpatrick's framework is comprised of four logically ordered categories of measurable criteria namely: (1) participants satisfaction/reaction; (2) learning or knowledge and skill acquisition; (3) transfer of learning or knowledge and skills to task; and, (4) training cost and benefit evaluation (Schrock & Coscarelli, 2007). It is necessary to begin with a brief description of each of the Kirkpatrick's levels of training evaluation including the types of evidence, encompassed by each level as summarized by Kirkpatrick and Kirkpatrick (2006) and Schrock and Coscarelli (2007).

(1) Participants satisfaction/reaction – This is the first level at which a training program should be evaluated. Schrock and Coscarelli called this type of evidence as a measure of customer satisfaction. It entails evidence about how the participants felt about the training, their satisfaction, and what they thought about the training. In operationalization this is typically in the form of self-reports of the participants on their satisfaction and perception of how they fared during and after the training.

(2) Learning or knowledge and skill acquisition – This is the second level at which a training program should be evaluated and includes evidence about changes in attitudes, knowledge, and skills as a result of participating in the training. Evidence at this level is criterion-referenced measures of how much of the training materials the trainee acquired i.e. the

competencies, knowledge and skills, taken from the training and hence the extent to which prescribed goals and objectives of the training program were met.

(3) Transfer of learning or knowledge and skills to task – At this third level the focus of evaluation is behavioral. Measures at this level are designed to assess the extent which the presumably acquired knowledge and skills from the training transfers to performance of the task.

(4) Cost and benefit evaluation – This level entails evaluation of the potential gains/benefits that would accrue to the trainer as a result of executing or using the training which includes estimate of returns to their investment on the program.

For the purpose of this dissertation evaluation the assumption is that participants had no prior knowledge about the modified Angoff standard setting tasks. However, in studies where the latter assumption is not met pre-test measures would help to facilitate making the claim that the training did in fact result in acquisition of knowledge and skills. Two evidential frameworks from standard setting literature informed evidence provided in this dissertation, at Kirkpatrick's first three levels and were:

- Raymond and Reid's (2001) framework for evaluating training
- Kane's (2001) validity framework

The Raymond and Reid's framework (2001) for evaluating training methods and for assessing whether the participants are well trained is based on three measurable criteria. In a nutshell, the three criteria require that the judgments of a well-trained standard setting participant should be stable over occasions, consistent with the assumptions of the standard setting method, and reflect realistic expectations. A little bit of explanation of the meaning of these three criteria immediately follows:

- Stable over occasions – This implies for example in the Angoff standard setting method context, that if a participant estimates the relative frequency of correct response to an item in the first round of rating as .50 then the expectation is that the item should receive about the same relative frequency estimate in subsequent rounds
- Consistent with assumptions – this implies that the standard setting judgments of a well-trained participant should conform with assumptions of the standard setting method including, about the standard setting materials, participants, and activities.
- Reflective of Realistic Expectations – This implies that outcomes of a well-trained participant should be reasonable at least when compared with available knowledge. For instance, in the context of Angoff standard setting, the relative frequency estimates of the items should be in an acceptable range as identified based on available knowledge.

The Kane (2001) approach is based on building an argument for or against the intended uses and interpretations of cut scores. There are three types of evidence to gather for making this validity argument in support of cut scores namely:

- Procedural validity evidence – the procedural evidence involves collecting information about the procedures used in establishing the cut scores including the degree to which the standard setting method was clearly defined and properly implemented;
- Internal validity evidence - the internal validity evidence includes information that supports or refutes the consistency within and between participants judgments;
- External validity evidence - the external validity evidence entails relating the cut score to external criteria such as other measures of student performance (Wyse, 2009).

The external criteria for evaluating participants judgments in this dissertation was based on fourth grade students responses to the MEAP test in 2005. The model for student responses that facilitated best guess estimate of item difficulty and cut score outcomes for the Proficient PLD for evaluating the judgments of participants in this dissertation was the unidimensional Rasch model. The unidimensional dichotomously scored Rasch item response model (Rasch, 1960) posits a non-linear monotonic relationship (called an item characteristic curve) between probability of correct response to a test item with person's abilities, and items difficulties. Accordingly, the probability of correct response to a test item depends on the difficulty of an item and the ability of the individual interacting with it, and individuals have a 50:50 chance of responding correctly to items of difficulty equal to their ability. The Rasch model is expressed mathematically as follows:

$$P_i(\theta) = P\left(X_i = 1 \middle| \theta \right) = \frac{\exp(\theta - b_i)}{1 + \exp\left(\theta - b_i\right)}$$

where: θ represents individuals abilities,

b_i represents item difficulties,

$P_i(\theta)$ represents the probability that an individual of ability θ responds correctly to the ith item of difficulty $b_{\text{i}.}$

5.2. Probability Evaluation Criteria

This section briefly reviews the frameworks for evaluating probability judgments. Also reviewed is the evaluation and analytic frameworks adopted for this dissertation, along with associated arguments and rationales.

It is important to point out that another identified problem with research on the Angoff method is that on the one hand, contemporary researchers assume implicitly or explicitly that participants probability judgments are subjective and introduce iterative feedback rounds of judgment that allow revisions of these probabilities in the light of new information. On the other hand, evaluation of these studies is virtually based on the frequentist interpretation of probability. The apparent mismatch between contemporary Angoff method standard setting research assumption of subjective probabilities and the frequentist approach to evaluation of results may have contributed to the gloomy view about the participant's capacity to perform the task. Therefore to illustrate this discrepancy between contemporary Angoff method subjective probability assumption and the frequentist approach to evaluation, it was deemed necessary to devote this section to clarification of analytic frameworks for evaluating probability judgments.

There are two identified approaches in the cognitive psychology literature to appraising the quality of probability judgments namely: coherence and calibration. The coherence and calibration probability evaluation frameworks have high generality. They have been applied in many practical settings as diverse as business, economics, political science, social policy, the law and medicine, etc., that require experts to make judgments under uncertainty and to produce subjective probability judgments (Arkes & Hammond, 1986; Bolger & Wright, 1993).

The coherence approach is based on appraisal of the quality of probability judgments by the degree to which judgments are consistent with each other and with the laws of probability theory (Keren, 1991; Yates, 1990). The calibration approach appraises probability judgment based on the criteria of extent to which the judgments correspond to the relative frequencies of the events to which they refer to (Bolger & Wright, 1993). The calibration approach typically employs plotting the proportion correct scores of participants based on objective measures of their performance against their assessed probabilities which is called the calibration curve. Based on the calibration curve the robust finding is that peoples' probability judgment are almost always monotonic with respect to the probability been assessed however their judgments tend to show over confidence and under confidence phenomena (Keren, 1991).

The measures for appraising probability judgments differ in degree of strictness. The strictness of a measure for appraising probability judgment is important because the graininess of evaluation criterion will affect the likelihood of observing skilled performance (Bolger & Wright, 1994). Precisely, if a strict criterion is used then fewer probability judgments will pass the test thereby leading to unfavorable conclusions about probability judgment than compared to if a weaker criterion was used. For example, a fairly loose coherence criterion require that a set of probabilities obey the principle of transitivity and a stricter requirement is that the probability conform to one or more of the four axioms of probability or to the Bayes' theorem (Bolger & Wright, 1993). On the other hand, a weak calibration criterion require monotonic relationship between judged relative frequencies with the true relative frequencies, while a stronger calibration criterion would require probability judgments to correspond to the true relative frequencies of the events to which they refer to (Keren, 1991).

The coherence and calibration criteria are also referred to in different areas of research using different terminologies. In the educational measurement literature the distinction is made between reliability and validity, respectively. Yates (1982) makes the distinction between internal consistency and external correspondence. Internal consistency refers to the importance of probability being reliable while external correspondence pertains to importance of probability judgments relating with the actual outcomes they refer to. Both terms also have been referred to respectively as the "syntactic" and "semantic" criteria respectively with the former meaning conformity to the algebra of probability and the latter implying the meanings of probabilities in the world. Winker and Murphy (1968) refer to these two standards as normative goodness and substantive goodness, respectively.

Normative goodness referred to the degree to which probability judgments truly reflects the judges belief and obey the axioms of probability theory while substantive goodness referred to the quality of the judge's knowledge of the domain of which the probability judgment is being made. The normative standard of goodness requires probabilities to correspond to judgments while the substantive standard of goodness requires probabilities to correspond to something in reality (Winkle & Murphy, 1968). Interpreting Winkler and Murphy's normative and substantive standards, Bolger and Wright (1993) inferred that coherence measures are indicators of probability knowledge while calibration measures are indicators of substantive domain expertise. Bolger and Wright (1993) also inferred that both domain knowledge and knowledge of probability axioms affects the coherence and calibration of probability judgment so that probability judgments should be assessed for both coherence and calibration. It was deemed necessary to highlight these equivalent dual probability judgment evaluation concepts because they all refer to the same thing it is just a matter of difference of terminology. However, because the concept of coherence and calibration are specific to the appraisal of probability judgments and for the sake of consistency with the educational measurement literature, the evaluation concepts of reliability and validity would be maintained for the purpose of this dissertation. The rest of the discussion in this section focuses on probability evaluation approaches in the standard setting literature.

There are three frameworks for evaluating the quality of Angoff method probability judgment outcomes. The three frameworks are: (1) Kane's (1994, 2001) validity framework; (2) Engelhard's (Engelhard & Anderson, 1998) Rasch model framework; and, (3) Reckase, (2006) and Wyse's (2009) psychometric theory. The Wyse (2009) approach was based on extension of the Reckase's (2006) psychometric framework. A little bit of background of these approaches to help establish the context for the approach adopted for evaluation of probability judgments in this dissertation follows.

Kane's (2001) approach based on validity criteria focuses on building an argument for or against the intended uses and interpretations of cut scores just as done with test scores. As specified in the last section, Kane's framework is based on three types of evidence in support of cut score validity namely: procedural, internal, and external validity evidence; Englehard's framework (Caines & Engelhard, 2009; Engelhard, 2007, in press; Engelhard & Anderson, 1998; Engelhard & Stone, 1998) applies the multifaceted Rasch model (MRM) to the probability judgments, while the psychometric approach (e.g., Reckase's, 2006; Wyse, 2009) is based on the assumption of a hypothetical intended cut score and is applicable with all unidimensional item response models, i.e. models of probability of a correct response to test items as a function of a single examinee ability and the item properties.

Kane's approach is by far the most common framework for evaluating the quality of standard setting results (Wyse, 2009). Of all three frameworks only Kane's approach is based on pure validation criteria. As noted by McGinty (2005) most of the existing standard setting outcomes evaluation frameworks are based on the conceptual umbrella of reliability. The notion of reliability also underlies the aforementioned Englehard's, Reckase's, and Wyse's psychometric evaluation frameworks. Also, Englehard's, Reckase's and Wyse's criteria are consistent with the predominant psychological theories of human rationality. The predominant psychological theories of human rationality are predicated on the notions of consistency, not of substance (Shafir & LeBoeuf, 2002). To elaborate, the criteria of rationality encapsulated in these approaches allow for flexibility of probability judgments provided they cohere in a normatively defensible fashion but not on substantive meaningfulness of the human judgments. It is however notable that computational as well as time, attention, memory, and similar limitations necessitate failures of ideal human rationality. Therefore, according to Shafir and LeBoeuf (2002), the aforementioned limitations to ideal human rationality implicates that the idea of human rationality remains at least in some sense intuitive rather than purely technical in nature.

This dissertation adopts a substantive focus to evaluation of probability judgments. It is based on extension of Kane's validity criteria. Hence, in addition to evidence about reactions of participants to training and pertaining to execution of procedures, this dissertation provides also, evidence about the reasonableness of cognitive processes, probability judgments, and cut scores. Specifically, evidence provided encompasses assessment of cognitive processes, substantive content domain knowledge, and correspondence of participant judgments with the empirical relative frequencies of the event of correct response to test items.

Also, besides the conventional correlational indices that are used for evaluating validity of standard setting outcomes with the Kane's validity criteria, this dissertation data analysis validation effort includes the principal coordinate's analytic (PCOA) technique that is based on Euclidean distance indices. The Non-Metric multidimensional scaling approach that is based on less stringent monotonicity assumption would have been more appropriate given the Heuristic training instructions tailored to the feature matching theory of similarity judgment (Tversky, 1977). However the rationale for the PCOA approach which like the correlational approaches is based on the more stringent linearity assumption is also to allow testing for the conventional assumption that a few dimensions underlie standard setting judgments.

The fundamental rationale for adoption of the comprehensive validation approach that includes the PCOA is because, it maps well with the Heuristic training for the probability judgment and also to allow for adequate evaluation of the judgment data for fit with the heuristic model principles. Also, because the Heuristic training method does not incorporate training on probability axioms and principles, it implies that evaluation of probability judgments for domain knowledge is more appropriate than assessment of probability knowledge. Besides, it is fair to hold participants accountable only for what they were taught, and only then is it meaningful to interpret and qualify their performance.

The decision to adopt a validation criterion for this dissertation was guided by Gigerenzer et al. (1999) work that highlighted that the function of heuristics is not to be coherent rather it is to make reasonable, adaptive inferences about real social and physical world given limited time and knowledge. Accordingly, the validity criterion levels the playing field for all cognitive strategies and makes less stringent assumptions about human judgments. For instance, the validation approach to evaluation differentiates logic from adaptive behavior and assumes that compliance with formal logic and probability principles does not necessarily imply high level of accuracy (Gigerenzer et al., 1999). According to the proponents of the adopted evaluation approach, although the heuristic may violate the conjunction, additive, and transitive coherence principles, they nevertheless may make fairly robust and accurate inferences.

5.3. Statistical Methods

The subsections of this section delineate the statistical methods that were used for this dissertation data analysis. The methods discussed includes those used for testing the plausibility of the assumptions delineated in Chapter four, the conceptual framework of this dissertation, and for addressing the research questions. Because of the substantive focus of data analysis in this dissertation, all descriptions of statistical methods in this section are in non-mathematical terms.

5.3.1. The Principal Coordinates Analysis

There are two statistical frameworks for exploring patterns in the data namely: person-centered and variable centered methods. Person-centered analytic methods focus on experimental participants, their relationships and interaction with tasks and make less restrictive assumptions about data structure. An example of a person-centered method is the Principal Coordinate Analysis (PCOA). On the other hand, variable-centered methods focus on experimental tasks, especially relationships between them, and make a number of restrictive assumptions about data (such as independence of observations). An example of variable-centered method is the Principal Component Analysis (PCA).

Although indices of correlation would also be obtained for this dissertation data analysis however, it is important to highlight that the person-centered analytic method of Principal Coordinate Analysis (PCOA) based on Euclidean distance indices is the preferred analytic framework for exploring the qualitative features of the standard setting data. The reason for preference of the PCOA method for the purpose of understanding the standard setting data are as follows: (1) to allow testing the conventional assumption that a few dimensions underlie judgments; (2) to recover new meaningful underlying variables that describe the data and to facilitate understanding of the data generating process (Webb, 2002); and, (3) standard setting data violates the assumptions of variable centered methods. To mention but a few features of standard setting data that implicate violation of assumptions of variable centered methods:

- Standard-setting studies often involve non random sample of participants limiting use of parametric inferential procedures. Put differently the data are often obtained from convenience samples not generated from a known probability mechanism such as random sampling.

- The number of participants used in standard setting studies is often much less than the number of variables leading to linear dependencies i.e. data set tends to be small and collinear.
- The assumption of statistical independence is violated because the participants are allowed to discuss as part of the process leading to complex dependency structures in the emerging data.

The Principal coordinates analysis (PCOA) is a technique sometimes referred to as geometric or ordination method that is based on analyses of matrix of distances or dissimilarities (the proximity matrix). It is used for representing data in a reduced dimensional space (Borg & Groenen, 1997; Webb, 2002). It differs from variable centered methods such as principal components and factor analyses techniques which operate on correlation matrices or angles between vectors. PCOA procedure systematize data, smoothes out noise, and provides graphical displays representing the similarities of objects and general structure of data that is much easier to understand than an array of numbers (Borg & Groenen, 1997; Schiffman, Reynolds, & Young, 1981). Classifying and organizing concepts are essential because they facilitate systematizing large amounts of data and aids human understanding. Hence, PCOA procedure would serve to help systematize the Angoff standard setting research data where organizing concepts and underlying dimensions are not well developed.

In formal terms, the problem of PCOA is as follows: Given a data matrix of distance measures between a set of objects, to find coordinates of the objects in a lower dimensional space so that the distance between a pair of objects is as close as possible to the their distances in the original space. Although the procedure makes no assumption about the existence of clusters in the data**,** they represent objects judged experimentally similar to one another as points close to each other and objects judged to be dissimilar as points distant from one another in a resultant spatial map (Schiffman, Reynolds, & Young, 1981)**.** The term “object” is used in a general sense to include humans. In statistical language, the procedures entail transformation of the original data using all variables to a data set with a reduced number of variables. Thus in executing these procedures all available variables are used and the data are transformed using linear or nonlinear transformation to a reduced dimension space. For the PCOA the objective function optimized is that measuring the discrepancy between the given dissimilarities and the derived distances. PCOA assumes that data are quantitative and therefore derives a functional relationship between the inter-point distances.

5.3.2. Bootstrapping

The non-parametric method applied in this dissertation study is the bootstrap resampling method (Efron, 1979; Good, 2006; Mooney & Duval, 1993). The bootstrap resampling method was used to operationalize the Proficient performance PLD. Specifically, it was applied to estimate hypothetical Proficient cut score for the Practice and the Real tests based on criterion referencing to the knowledge and skills descriptors of the PLD. The estimated cut scores for the tests were used for the purpose of cross validating the estimates based on the Heuristic and Normative training participant’s judgments.

The bootstrap resampling technique involves drawing with replacement samples from the original random sample taken from a population. The basic assumption in bootstrap method is that the original sample is representative of the population. If the latter holds then one can mimic sampling from the population by sampling from the sample. The bootstrap method is useful for estimating the sampling distribution of a statistic, including its standard error, bias, and for forming confidence interval for the underlying parameters. However, validity of bootstrap estimates depends on the quality of the original sample. There are a number of techniques for finding bootstrap confidence intervals with the percentile method being the simplest. The percentile method works well, when the bootstrap distribution of the statistic is symmetrical and centered around the parameter.

5.3.3. Statistical Indices

The statistical indices computed for evaluating the comparative technical qualities of reliability and validity of the judgment outcomes of the training methods were as follows:

- Cut scores – the cut score for each participant was obtained as the sum of their item level probability judgments and represent the mean of a generalized binomial distribution, also called true score.
- Means and standard deviations – three different estimates of item level mean were used namely: empirical item means of fourth grade student population, study group means, and bootstrap PLD item mean estimate. Details of the PLD bootstrap mean estimate is provided in the results section.
- Correlations – these are numerical summaries of bivariate relationships (Huck, 2004).Two types of bivariate correlations representing the different scales of measurement were computed in this dissertation and were: Spearman Rho and Pearson product moment correlation. The Spearman Rho is the appropriate correlation index when the variables are measured in an ordinal or rank ordered scale and is based on the assumption of monotonicity while the Pearson correlation is appropriate when variables are measured on an interval scale and the assumption of linearity is met.

5.3.4. Statistical Inference Methods

Two statistical inference techniques were used to compare the Heuristic and the Normative training judgment on the derived indices of correlations and cut scores for significance of difference and are: the Mantel test and the independent sample *t*-tests. The Mantel test was named after Mantel (1967) who proposed it. It is a non-parametric permutation test useful for computing and testing the significance of the differences of the correlation of the matrices of the same rank, of the distances between a set of objects. The Mantel test of significance was used to evaluate the reliability of the judgments of the Heuristic and the Normative training participants. For the purpose of this dissertation, the Mantel test was based on correlating and testing significance of the difference from 0 of the correlation between the Euclidean distance matrices of the rounds of judgment data of the items that were replicated on both the Practice and the Real tests. The independent sample *t*-tests were used to compare the means on the derived outcome indices for the Heuristic and the Normative training groups.

Chapter Six: Methods

This chapter is comprised of four major sub-sections that delineate characteristics of the participants, stimuli, design and design variables, and procedures employed for the empirical studies. Two standard setting studies were run, one based on the principles of the cognitive heuristic model called the Heuristic training while that based on typical elements of the Angoff methods is the Normative training.

6.1. Participants

There were 10 and 12 participants in the Heuristic and the Normative training, respectively. The participants were recruited through e-mails sent out via list serves and through classroom visits. The participants were comprised mostly of Michigan State University (MSU) Teacher Education (TE) pre-service teachers and teachers in the mid-Michigan area.

Due to scheduling conflicts, 18 out of 22 of the participants chose the date to participate in the study albeit they did not know the type of training they would be receiving. Consequently, the assignment of participants to training group could still be considered as approximating a random process. Meanwhile, the remaining four participants that indicated they could participate on both dates were assigned to a training group at the discretion of the researcher, mainly in consideration of balance of group size also, because one of the two scheduled dates was significantly less chosen.

In terms of demographics, the participants were mostly white females. There were 21 white (9 out of 10 participants in the Heuristic and all 12 participants in the Normative training), and one Asian (in the Heuristic training). There were five males (3 out of 10 participants in the Heuristic and 2 out of 12 participants in the Normative training), and 17 females (7 out of 10 participants in the Heuristic and 10 out of 12 participants in the Normative training). The areas of specialization that the participants represented were social studies, kinesiology and cognitive neuroscience, teacher education, educational policy, and mathematics. About a half of the participants in both groups were mathematics specialists. There were 13 mathematics (5 out of 10 participants in the Heuristic and 8 out of 12 participants in the Normative training), three social science (1 out of 10 participants in the Heuristic and 2 out of 12 participants in the Normative training), two teacher education (1 in each of the training groups), three educational policy (all in the Heuristic training), and one kinesiology and cognitive neuroscience (in the Normative training) specialists.

More than a half of the participants in both training had teaching experience and mostly in urban school districts. To further describe the participants, 17 of the participants had teaching experience (8 out of 10 participants in the Heuristic and 9 out of 12 participants in the Normative training), five had no experience teaching (2 out of 10 participants in the Heuristic and 3 out of 12 participants in the Normative training). The three school district locales (urban, rural, and suburban) were represented in the studies and, of the teachers that responded to the question of the school district they had taught, there were 11 urban (4 out of 10 participants in the Heuristic and 7 out of 12 participants in the Normative training), three suburban (2 out of 10 participants in the Heuristic and 1 out of 12 participants in the Normative training), one rural (in the Normative training). The grade levels at which they had taught were as follows: six had taught at the kindergarten through grade 2 level (2 out of 10 in the Heuristic and 4 out of 12 in the Normative training), eight had taught at grades 3 through grade 5 level (5 out of 10 in the Heuristic and 3 out of 12 in the Normative training), nine had taught at the grades 6 through 9 level (4 out of 10 in the Heuristic and 5 out of 12 in the Normative training), four had taught in the grades 9 through 12 levels (1 out of 10 in the Heuristic and 3 out of 12 in the Normative training). More than a half of the participants had taught mathematics specifically, 14 had taught mathematics (seven in each group), and eight had not taught mathematics (3 out of 10 in the Heuristic and 5 out of 12 in the Normative group).

The current positions held by the participants were as follows: five were currently K-12 teachers (3 out of 10 in the Heuristic and 2 out of 12 in the Normative training), five were Michigan State University(MSU) doctoral students (3 out of 10 in the Heuristic and 2 out of 12 in the Normative training), two were masters level teachers (one in each training group), seven were Michigan State University (MSU) Teacher Education undergraduate pre-service teachers (2 out of 10 in the Heuristic and 5 out of 12 in the Normative training), one was a K-2 assistant principal (in the Heuristic training), one was a curriculum specialist (in the Normative training), and one participant was currently not in any educational field (in the Normative training).

The non-cognitive attributes of the participants that were measured were their intrinsic and extrinsic motivation, emotion, and, engagement. Each of these constructs was measured by 4-point Likert scale items. The items were adapted from an on-going Research and Evaluation on Education in Science and Engineering (REESE), National Science Foundation (NSF) funded project that the researcher was a part of. The principal investigator of the REESE NSF project is David Kantor under the auspices of the New York Hall of Science (NYSCI). The scales from which the items were adapted from have been pilot tested and all had reliability above .70's. Both groups scored above average on the measures of these constructs and showed higher average intrinsic than extrinsic motivation.

6.2. Stimuli

The stimuli used for both studies were the Proficient performance level descriptor (PLD) and tests. These are described in turn in the sub-sections that immediately follow.

6.2.1. The Performance Level Descriptors (PLDs)

The Michigan Educational Assessment Program (MEAP) fourth grade mathematics Proficient performance level PLDs that was developed in 2005 was used for both studies. Cut score was sought for the Proficient level of performance. Although the PLD was for the fourth grade, the grade level content expectations (GLCE) were for the third grade mathematics content because the MEAP test is administered in the fall of each year over skills that were taught the previous year. Meanwhile, each PLD addressed at least one of the third grade level content expectations. Table 6-1 shows the Proficient PLD that was used to facilitate both studies.

Table 6-1: The 2005 Fourth-Grade MEAP Proficient PLD

Proficient PLD
Read, write, and compare whole numbers up to 10,000.
Fluently solve and estimate basic problems using addition and subtraction with two-digit numbers with regrouping and up to four-digit numbers without regrouping. Solve and apply basic multiplication and division problems up to 10x10.
Recognize, name, and solve problems involving common fractions and decimals including money.
Calculate, apply, and use common units of measure in length, weight, time, and temperature in contextual situations.
Identify, describe, compare, manipulate, and construct common two- and three-dimensional objects. Identify properties of lines.
Read, interpret, and solve problems involving bar graphs.

6.2.2. Tests: The Practice and the Real Tests

The content area of the tests was mathematics. The conceptual framework of the knowledge and skills measured by the tests was provided by the Michigan Curriculum framework (MCF) and the Webb’s Depth of Knowledge (DOK) levels (http://www.michigan.gov/mde). The MCF is the conceptual framework for the content while the Webb’s DOK is the conceptual framework for the cognitive processes measured by the mathematics test items. It was considered appropriate to provide brief descriptions of the MCF and Webb’s DOK before giving the specifics of the Practice and the Real tests. The MCF was developed by the Michigan Department of Education (MDE) in 2004 in response to the *No Child Left Behind Act* (NCLB). Math content was organized at three levels of hierarchy across grade levels in the MCF. At the top of the hierarchy are the content strands, at the middle are the domains, and at the base are the grade level expectations. There are multiple domains in each content strand and several expectations within each of the domains. The Webb's depth of knowledge (DOK) levels posits the cognitive processes required in responding correctly to the mathematics test items. There are four DOK levels, designated in ascending order of cognitive complexity as: recall (RE), skills and concepts (SC), strategic thinking (ST), and extended thinking (ET). The 2005 MEAP test items were written to target the DOK levels. Table 6-2 presents the structure of the MCF.

Table 6-2: Michigan Mathematics Curriculum Organizational Structure

Number & Operations	Measurement	Geometry	Data and Probability
Domains			
Meaning, notation, place value, and comparisons (ME)	Units and systems of measurement (UN)	Geometric shape, properties, and mathematical arguments (GS)	Data representation (RE)
Number relationships and meaning of operations (MR)	Techniques and formulas for measurement (TE)	Location and spatial relationships (LO)	Data interpretation and analysis (AN)
Fluency with operations and estimation (FL)	Problem solving involving measurement (PS)	Spatial reasoning and geometric modeling (SR)	Probability (PR)
		Transformation and Symmetry (TR)	

Table 6-2 was adapted from the MEAP website (http://www.michigan.gov/mde).

Now to the specifics of the tests, the tests used for practice and the two rounds of Angoff method judgments were referred to as the Practice and the Real test, respectively. Both were comprised of 15 multiple choice items. They were subsets of released 2005 and 2006 fourth grade MEAP test items that were developed in accordance with the Michigan Curriculum Framework. The four content strands and 13 third grade expectations were represented on both tests. The content strands represented were: Number and Operations (N), Geometry (G), Data and Probability (D), and Measurement (M). Items designated at Level 1 (recall) and level 2 (skills and concepts) of the Webb's depth of knowledge (DOK) levels, were represented on both tests.

In order to create adequate content matched pairs, 25 items were selected from the 2005 and five items were selected from the 2006 released MEAP tests. Statistical information was unavailable for the 2006 items (two were on the Practice test and three items on the Real test).

As a consequence, statistics for the matching 2005 pair were substituted for the 2006 items. All the same, items were carefully selected to represent and align well with: the knowledge and skills of the Proficient PLD, the content strands assessed by the MEAP test and, for which content experts had considerable agreement about the DOK, and with adequate mean square fit indices to the Rasch item response model.

The Practice and the Real test were matched in terms of content (i.e., content strands, content domain and, grade level content expectations) but differed in terms of DOK level of items. The Practice and the Real test booklets were assembled so that items of equivalent grade expectation appeared in the same position on both booklets. Five items were replicated on both booklets and appeared in the same positions on both. There were more items of DOK level 2 in the Real test and precisely, it was meant to be more difficult, with the rationale being to facilitate evidence about substantive meaningfulness of feedback rounds of judgments of the participants. The expectation was that the group discussion polarization phenomenon might show up, so that to the extent that participants' judgments on the Real test generate higher cut score implies normative influence, and inappropriate influence of feedback. The aforementioned test matching criterion features (content strands, grade level expectation, and DOK) were highlighted also because they were essential cues for the judgment task of estimating relative frequencies of correct response on the items. Tables 6-3 and 6-4 that follow shows features of the tests including: assembling design, GLCE, and the DOK levels.

Table 6-3 shows the Practice and the Real test assembling design including: item position, indictor of whether item was replicated on both, content strands, domains, grade level content expectations codes, and depth of knowledge level of the item that appeared in each position for each of the tests. Please note that the DOK designated to items were the modal DOK assigned to the items by 13 content experts that participated in the 2005 MEAP test alignment study. Table 6-4 contains descriptions of the grade level content expectations measured by the Practice and the Real test and for which corresponding GLCE codes are specified in Table 6-3.

Table 6-3: The Practice and the Real Tests Design

Position	Replicated	Content Strand Code	Domain Code	Grade level Expectation Code	Practice Test Item DOK Level	Real Test Item DOK Level
1	No	N	ME	N.ME.03.02	1	2
2	Yes	N	MR	N.MR.03.10	2	2
3	No	N	ME	N.ME.03.01	1	1
4	No	N	ME	N.ME.03.02	1	1
5	No	N	FL	N.FL.03.06	1	1
6	Yes	N	ME	N.ME.03.03	2	2
7	No	N	ME	N.ME.03.16	1	1
8	Yes	M	UN	M.UN.03.02	2	2
9	No	M	UN	M.UN.03.04	1	1
10	No	G	GS	G.GS.03.06	1	2
11	Yes	D	RE	D.RE.03.02	2	2
12	Yes	G	SR	G.SR.03.05	1	1
13	No	D	RE	D.RE.03.03	2	2
14	No	M	UN	M.UN.03.02	2	2
15	No	M	UN	M.UN.03.02	2	2

Notes:
Content Strand Codes: N (Number and Operations); M (Measurement); G (Geometry); D (Data and Probability); Domain Codes: ME (Meaning, notation, place value, and comparisons); MR (Number relationships and meaning of operations); FL (Fluency with operations and estimation); UN (Units and systems of measurement); GS (Geometric shape, properties, and mathematical arguments); SR (Spatial reasoning and geometric modeling); RE (Data representation); The Grade Level Expectation is coded with a strand, domain, grade-level, and expectation number. For example, N.ME.03.02 indicates: N - Number and Operations strand; ME - Meaning, notation, place value and comparison domain of the Number and Operations strand; 03 - Grade 3 Expectation; 02- Second Expectation in the third Grade-Level of the ME domain and the Number and Operations strand

Table 6-4: Third Grade Content Expectations Measured by the Tests

Grade Level Expectation	Description
1	Read and write numbers to 10,000
2	Add and subtract fluently two numbers; up to and including two-digit numbers with regrouping and up to four-digit numbers without regrouping
3	Compare and order numbers up to 10,000
4	Recognize multiplication and division situations
5	Understand that fractions may represent a portion of a whole unit that has been partitioned into parts of equal area or length
6	Measure in mixed units within measurement system
7	Understand sample decimal fractions in relation to money
8	Know benchmark temperatures and freezing
9	Identify, describe, compare and classify three-dimensional solids e.g. prism
10	Read scales on axes, Identify the max, min, range
11	Compose and decompose triangles and rectangles to form other familiar two-dimensional shapes e.g. form a rectangle using two congruent right triangles, or decompose a parallelogram into a rectangle and two right triangles
12	Solve problems using information in bar graphs, including comparison of bar graphs
13	Identify place value of digit in a number

6.3. The Design of Empirical Study

6.3.1. The Design

The intended design of the study is two-way mixed effects Analysis of Variance (ANOVA). In formal language a 2×3 mixed effects ANOVA. However, because of scheduling conflict, the participants were allowed the option of choosing the date they would participate in the study. All the same, one could still consider the training method assignment process as random because the participants were unaware of the type of training that they would be receiving. Meanwhile, a few of the participants that indicated that they could participate on any of the dates were assigned at the discretion of the researcher to one of the training dates to create the required balance in gender, teaching experience, and group sizes.

As presented in Table 6-5, training method is a between factor while training round is a within factor. The main factor is training method which is comprised of two levels (i.e. Heuristic and Normative training). There were three training rounds of judgments within each training method. For both studies, instructional and practice activities preceded the first round of judgment while the second and third rounds of judgment followed feedback to the participants on their preceding round of judgment. The training interventions of instruction, practice, and feedback for the Heuristic training were tailored to the heuristic model principles while those for the Normative training modeled the prevailing Angoff training methods. Table 6-5 shows the design of the study.

Table 6-5: The Design of the Study

Training Method	Practice Round	Feedback Round One	Feedback Round Two
Heuristic			
Normative			

Participants were split into table groups within each training method. There were seven table groups in all, three table groups in the Heuristic training (two tables with three participants each, and one table with four participants), and, four table groups in the Normative training (four tables with three participants in each).The background characteristics explicitly considered in assigning the participants to table groups were the indicator of whether they had teaching experience, their current position, and their gender. For instance, it was ensured that at least one participant with teaching experience was assigned to each table group, and that the minority male gender sat on different table groups. The rationale for assigning participants to table groups so that at least one participant with teaching experience was on each table was also to augment for deficiencies in firsthand empirical knowledge of students through the performance level description training discussion exercise. Tables 6-6 and 6-7 provides table representation of the variables explicitly considered for allocating the Heuristic and the Normative training participants, respectively to table groups for discussion.

Other potential table group allocation background variables that were not used in this dissertation for table group allocation design included: number of years teaching, number of years in educational field, indicators of math specialization, experience teaching mathematics and at grade level of interest for standard setting. The reason why these variables were not considered in the table group allocation design was because they were unknown prior to the study. The participants were allocated to table groups prior to the day of the study while they turned in survey about their backgrounds on the day of the study.

Table 6-6: Heuristic Training Table Group Allocation Variables Distribution

Table Group	Gender	Taught	Current Position
1	Male	Yes	Teacher
1	Female	Yes	Teacher
1	Female	Yes	Assistant principal
2	Female	Yes	Teacher education doctoral student
2	Female	No	Educational policy doctoral student
2	Female	Yes	Teacher
2	Male	Yes	Pre-service teacher education senior
3	Male	Yes	Teacher education doctoral student
3	Female	No	Pre-service teacher education senior
3	Female	Yes	Teacher

Table 6-7: Normative Training Table Group Allocation Variables Distribution

Table Group	Gender	Taught	Current Position
1	Female	Yes	Teacher
1	Female	Yes	Pre-Service teacher education senior
1	Female	Yes	Math curriculum specialist
2	Male	Yes	Teacher
2	Female	Yes	Doctoral student curriculum teaching & educational policy
2	Female	No	Pre-service teacher education senior
3	Female	Yes	Doctoral student kinesiology and cognitive neuroscience
3	Female	No	Not currently in education
3	Male	No	Pre-service teacher education senior
4	Female	Yes	Teacher
4	Female	Yes	Pre-service teacher education senior
4	Female	Yes	Pre-service teacher education senior

6.3.2. The Design Variables

Extraneous variables that could potentially impact the outcome of the study explicitly considered and measured by the design included non-cognitive constructs namely: motivation, engagement, and emotion. Also, the background characteristics that were measured that could potentially confound with the study outcomes includes: number of years in teaching, number of years in educational field, indicators of math specialization, experience teaching mathematics, and at grade level of interest for standard setting.

Teaching experience was measured by an open ended item that asked participants about the number of years they had been teaching. The teaching experience variable had missing data because the researcher omitted asking those that were currently teaching about how many years they had been teaching. Educational experience was measured by an open ended item that asked participants about the number of years they had been in the educational field. The constructs, motivation, engagement, and emotion were measured by composite, sum scores of 4-point Likert items with response options strongly disagree to strongly agree.

The independent variable was training method with two levels (the Heuristic and the Normative training). As you will see later in the study procedures section, the Heuristic and the Normative training differed on the dimensions of instruction, practice, and feedback. The stimuli factors considered by the design of the training were the Proficient PLD and constructs measured by the test items namely: content strands, GLCE, and DOK levels.

The primary dependent variable was participants judged conditional probabilities that the barely proficient fourth graders would respond correctly to the Practice and the Real test items. These judged conditional probabilities are also referred to as item difficulties for the barely proficient students or conditional item means. However, for the remaining discussion in this section, the specialized measurement terminology of item difficulties would be used.

Item difficulties were judged by participants on the percentage scale (0-100). These were then converted to proportion correct scale for analysis. The judgments were in response to the question of the percentage of students who are barely proficient that would respond correctly to the items on the test. The internal criterion variable used for evaluating participants judgments were the pooled study group mean item difficulty estimates, while the external criteria variables were the entire fourth graders empirical proportions of correct response to the items in 2005, modal DOK designations of item by content experts that participated in alignment study in 2005, and the bootstrap PLD cut score item difficulties estimates for the Practice and the Real test.

Derived dependent variables for comparing the training methods were the participants judged difficulty ranks of items, correlations of participants judged item difficulties with internal and external criterion variables, correlations of participants judged difficulty ranks of items with item difficulty ranks based on internal and external criteria, mantel correlations of distances between pairs of replicated items across rounds based on participants judged difficulties of the items, cut scores, and deviations of participants cut scores from internal and external criterion variables.

6.4. Empirical Study Procedures

The purpose of this section is to present the structural features of this dissertation's empirical studies and to draw out the contrasting features of the training methods. Specifics of procedures employed with each of the studies are provided in the Heuristic and the Normative training subsections which immediately follows. Figure 6-1 displays the structural features of both studies. A brief description of the information represented in Figure 6-1 is provided in their order of precedence in the immediately following paragraph.

Five days prior to each training session, a survey was sent out by e-mail to the participants that was labeled participant information sheet. The survey elicited participant's demographics, experience, and motivation. The participants were instructed to come in with filled out survey (please refer to the Appendix C for the participant information survey questions).

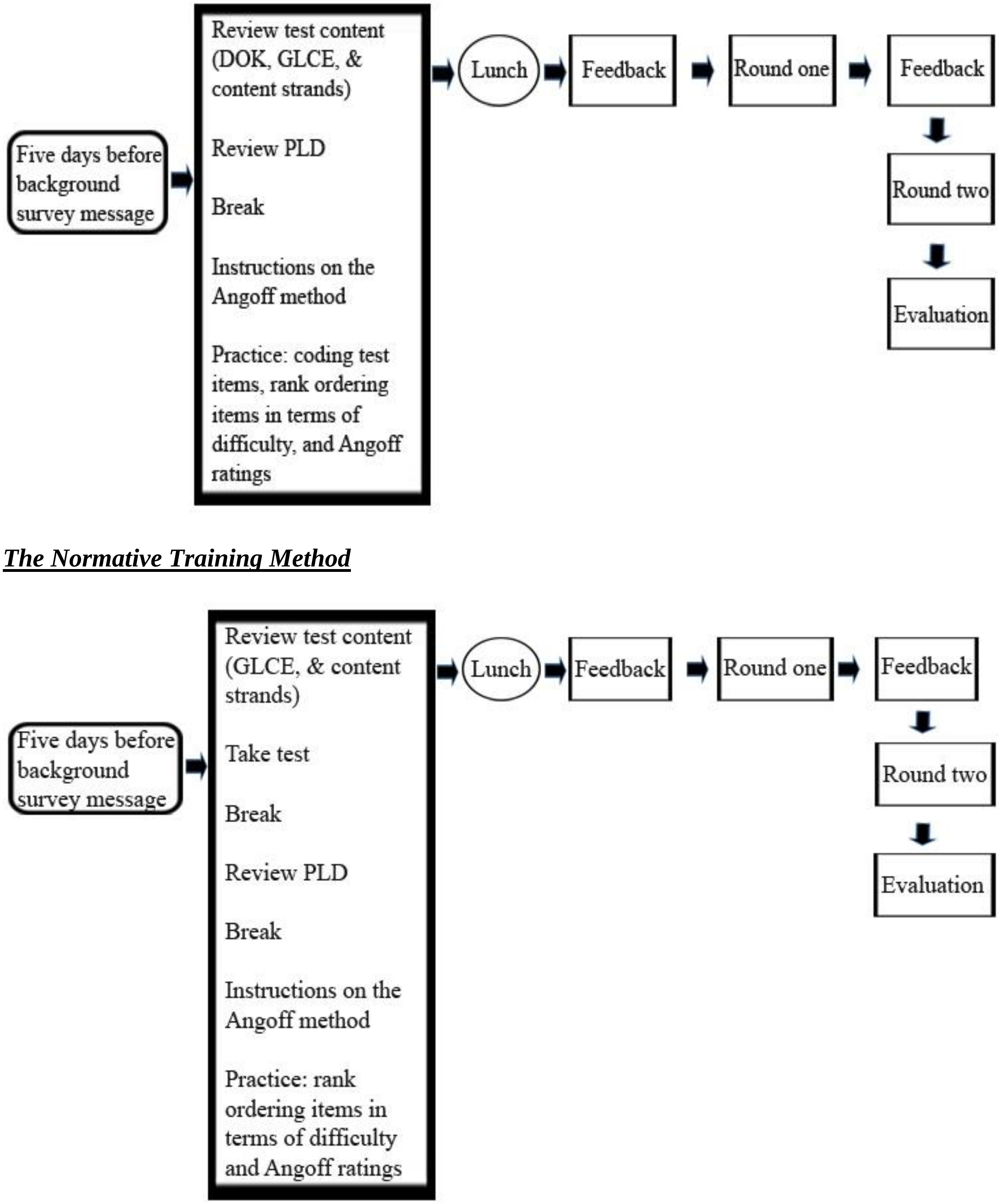
Both studies were conducted onsite at Michigan State University and were facilitated by the researcher assisted by an MSU educational policy expert, using detailed scripts written by the researcher. (See Appendix B for sample scripts.) Both sessions began with introductions, review of the purpose of study, and review of housekeeping matters (e.g., signing the institutional review consent form, and briefing about food and the honorarium of $100).

All training activities were timed and all questions addressed by participants during the training were handed out in survey format. (See Appendix C for surveys and practice exercises). The activities prior to and including the practice round of judgment for the Normative training was comprised of essential components of typical Angoff standard setting training such as the following: review of content domain background of the tests, taking the test, PLD review, instructions on the modified Angoff procedure, and practice judging difficulties of test items. The activities prior to and including practice round of judgment for the Heuristic training differed from that of the Normative training in terms of the following: instruction on Webb's Depth of Knowledge levels (DOK), explicit instructions on the heuristic strategy for rating items, and practice coding items to assessment categories. The practice round of judgment for both training groups was followed (after lunch), by feedback. In addition to the practice round of judgment there were two rounds of judgments, the Real test rounds of standard setting judgments, with feedback also provided between the two rounds and for both studies.

The studies were each concluded with the administration of evaluation survey. The evaluation survey elicited participant satisfaction, engagement, emotion, factors considered, perceived influence of feedback, understanding, and confidence in judgments. A substantial number of the evaluation questions were adapted from the Michigan Educational Assessment Program (MEAP) 2005-2006 Technical Manual, Cizek and Bunch (2007), Missouri end of course standard setting study, and the REESE NYSCI project (see appendix C for the evaluation questions).

Figure 6-1: Study Procedures

The Heuristic Training Method



The Normative Training Method

Five days before background survey message
Review test content (DOK, GLCE, & content strands)
Review PLD
Break
Instructions on the Angoff method
Practice: coding test items, rank ordering items in terms of difficulty, and Angoff ratings
Lunch
Feedback
Round one
Feedback
Round two
Evaluation
Five days before background survey message
Review test content (GLCE, & content strands)
Take test
Break
Review PLD
Break
Instructions on the Angoff method
Practice: rank ordering items in terms of difficulty and Angoff ratings
Lunch
Feedback
Round one
Feedback
Round two
Evaluation

Contrasting features of the Heuristic and the Normative training methods are presented in Table 6-8. The methods differ on the training dimensions of instruction, practice, and feedback. The Heuristic and the Normative training instruction differs in terms of explicitness about the judgment factors and on how to integrate them into judgment, while the practice exercise differs in terms of extent of content information processing of the test items. There were two feedback types used for the training as shown in Table 6-8: intermediate-outcomes and final-outcomes feedback. The Heuristic training received both intermediate-outcome and final-outcome types of feedback. The intermediate outcome feedback was the substantive domain item difficulty ordering (aka, DOK).The Normative training received more elaborate final-outcome feedback that included conditional probabilities of correct response at some student ability levels (aka, construct map feedback).

It is important to mention that the Heuristic and the Normative training studies were not contrasted on the dimension of timing of activities. Training activities were timed similarly for both studies. However, timing of activities was left out from descriptions of study procedures because it was not completely followed. Specifically, the participants finished tasks ahead of scheduled time.

Table 6-8: Contrasting the Heuristic and the Normative Training Methods

Factor	Heuristic	Normative
Instruction	• Explicit instruction on factors to consider in the judgments • Explicit Instruction on strategy to integrate factors into judgments • The factors specified were the content strands, GLCE, and DOK of items	• Implicit instruction on factors to consider in the judgments • The factors implicitly specified were the content strands and GLCE of items
Practice	• Code items to content strands, GLCE and DOK • Rank order items in terms of difficulty • Estimate proportion of fourth grade students who are barely proficient that would respond correctly to the items	• Take test • Rank order items in terms of difficulty • Estimate proportion of fourth grade students who are barely proficient that would respond correctly to the items
Feedback	Intermediate- outcomes feedback • Substantive domain item difficulty ordering- content expert DOK designations Final-outcomes feedback • Empirical proportions of the barely proficient responding correctly to the items • Empirical item difficulty ranks based on entire fourth grade students responses • Participants estimated item difficulties and cut scores exchanged through whole and table group discussions	Final outcomes feedback • Empirical proportions of the barely proficient responding correctly to the items • Empirical item difficulty ranks based on entire fourth grade students responses • Proportion of students responding correctly to the items at some ability levels called construct map feedback • Participants estimated item difficulties and cut scores exchanged through whole and table group discussions

6.4.1. Procedures for the Heuristic Training

The Heuristic training began with introductions led by the researcher. Name tags and Power Point slides were used to facilitate these introductions. The facilitators introduced themselves by indicating their names, affiliation, and area of expertise. After the facilitators introduced themselves, the participants were asked to introduce themselves by stating their name, background, why they agreed to participate in the meeting, and what they expect to get out of the meeting. The latter two questions were asked also in order to gauge participant's motivation.

Following introductions, the researcher briefed participants about some housekeeping matters, an essential part of which was completing the Institutional Review Board (IRB) consent form. After going through house-keeping matters, the co-facilitator took over leading training discussions. The introductory part of the training entailed review of the purpose and agenda for the training and these were projected on the overhead as well as read out from the script by the co-facilitator.

The first training activity was the review of background of the subset of the MEAP mathematics tests that were used for the study. For review of the background of the MEAP tests, the participants were given handouts delineating the content strands and GLCE's measured by the items on the Practice and the Real tests, definitional sheet for major assessment concepts referenced by the study, and definitional sheet for the Webb's depth of knowledge levels with an example of item designated at each level. The training instructions specified that the content area of the tests was mathematics, that the grade level of the tests was fourth grade, the items on the tests were multiple choice, there were 15 items on each of the tests selected from past MEAP tests, the content strands, GLCE's, and DOK levels measured by the tests. The instruction also specified that test items were pilot tested and reviewed using a rigorous process so that their quality was not in doubt. These instructions were summarized in bullet points and projected on the over head while reading them out from the script (refer to Appendix B for the script read out to the Heuristic training for review of the background of the Practice and the Real tests). At the end of review of the test content material, the participants were encouraged to ask questions before proceeding with the PLD review.

The second training activity was the review of the Proficient performance level descriptors (PLD). For the PLD review, each participant was given a hard copy of the Proficient PLD, and flip charts were handed out to each table group. The participants were instructed to elaborate on the PLD by giving three illustrations for each descriptor of what a barely proficient student should know and be able to. The instructions on how to elaborate on the Proficient PLD were projected and partly read out from the script (please note that the script was not completely followed for the PLD review). Following the table group discussions, each group presented their work during which the co-facilitator also led discussion to highlight similarities and differences in the table group descriptions of the barely proficient student. The group discussion continued until it was gauged that there was consensus among participants about the knowledge, skills, and abilities of the barely proficient students. Then the participants were allowed a short break.

The third training activity after the short break was review of the modified Angoff heuristic instructions. For this review, the instructions were projected, handed out as well as read out from the script (refer to the Appendix B for the complete modified Angoff heuristic instruction script). Meanwhile, the most essential part of instructions read out from the script on the Angoff method was as follows and italicized to distinguish it from the rest of the discussions:

The complete instructions to guide you with using this procedure to rate the items are contained in the modified Angoff rating instructions hand-out I just gave you, and are:

A. Think about the barely proficient students

For each item on the test:

B. Think about what it measures (Content strand, GLCE, and DOK level)

C. Think about items that measure these same knowledge and skills

D. Recall or imagine the proportion of students who are barely proficient that would respond correctly to items in this category.

E. Mark the percentage from 0 to 100

Any question? The overarching expectation is that you all rely on information provided in this training in giving your best judgment. However, for those of you with experience teaching fourth grade students, these instructions are really aimed at activating your knowledge of interaction of actual students in your classrooms, who match the descriptions of barely proficient with similar test items. Keep in mind that similarity of test items is defined for our purpose in terms of matching content strand, GLCE and DOK level. By all means, refer to the paper versions of these instructions and all materials provided in this training in making your judgments. Use your best judgment to make these decisions, but do not agonize over them.

The fourth training activity was the practice round of judgment. The practice round of judgment was based on the Practice test. Questions to address for the practice round of judgment were handed out to participants in survey format along with rating sheets and the Practice test booklets. The practice questions included in order: coding items to content strands, GLCE, and depth of knowledge levels, rank ordering items in terms of difficulty, and judging item difficulties for the barely proficient students (refer to Appendix C for the exact questions addressed by the Heuristic training participants in the practice round). For the practice exercises, the participants were instructed to read the directions thoroughly before addressing the questions, to ask for clarifications if they encountered stumbling blocks along the way, to turn in rating sheets when done, and to hold on to the Practice test booklets for feedback. When the participants completed the practice exercises, they took a lunch break. While the participants were at lunch their practice designations of items to DOK levels and judgments of the proportion of the barely proficient that would respond correctly to the items were analyzed. The data were analyzed by the researcher and assisted by a fellow student using Excel and SPSS spreadsheet packages. Bar graphs were generated showing for each of the items, how many of the participants designated it to each DOK level. Line charts were also generated for each item on the judgments of the proportion of the barely proficient that would respond correctly to it and for the cut score estimates of the participants. The cut score estimate of the participants were calculated as the sum of their item level judgments of the proportion of the barely proficient that would respond correctly to the test items.

The fifth training activity was feedback to the participants on the practice exercises. The participants were given feedback on their designations of the items to DOK levels and on their judgments of the proportion of the barely proficient that would respond correctly to each item. From this practice round feedback henceforth, the researcher took over facilitation of the study. The script was also followed for the feedback discussion. The practice feedback instruction emphasized that the goal was for the participants to compare judgments and share rationales and not to reach consensus. The practice feedback instruction also highlighted that the focus of discussion was on their DOK designations, judgments of the proportions of barely proficient that would respond correctly to the items, and cut scores because the content strands and the GLCE's are more objectively determined. The feedback summary statistics were projected on the overhead. For item level data summaries, three items were displayed per slide in order to make it easy for participants to digest the information. First, bar chart summaries were projected showing how many of participants designated each item to each DOK level. The discussion of the bar charts was process-oriented and structured to focus on items for which more than a half of the participants designated DOK that were discrepant from the modal DOK level assigned to the item by the 13 content experts that participated in the alignment study in 2005 (please note that the latter was the operational definition for the DOK of the items). Participants that designated DOK to items that were discrepant from those of the content experts were asked to share their rationale and the group deliberated on these rationales for a while before we proceeded with the next slide. This process was followed for all of such items. Second, line chart summaries were projected that contained on the horizontal axis the participants' ID and on the vertical axis, their judgments of the proportion of barely proficient that would respond correctly to the items. The line chart displays also showed summary statistics of participant's practice round judged proportions of the barely proficient that would respond correctly to the items (mean, standard deviation, minimum, maximum, and range of the judged proportions). Discussion of the line chart summaries of judged proportions of the barely proficient that would respond correctly to the items was also process-oriented and structured to focus on items for which there was greater discrepancies of the judged proportions. Participants with outlying judgments were asked to share their rationales and the whole group deliberated on these views for a while before proceeding with next graphical display. This process was carried through for all such items.

The third as well as the last data summary displayed was the line chart of the cut score estimates of the participants for the Practice test. The line chart contained on the horizontal axis the participants' ID and on the vertical the cut score estimates along with the associated group summary statistics (mean, standard deviation, minimum, maximum, and range). The same process of asking participants with outlying estimates to share their rationales was followed for deliberation on the cut scores. Participants were curious to know how the cut scores were derived, so as a consequence the researcher also discussed the cut score computation process. At the end of the feedback discussion, the participants were given a handout containing the content strand, GLCE, DOK, the empirical rank ordering and the proportions of the fourth grade students responding correctly to the Practice test items in 2005. After the practice round feedback, the Practice test booklets were collected from the participants and the standard setting activities based on the Real test booklet began.

The sixth training activity was the first round of standard setting judgments based on the Real test booklet. For this first round of judgment, the participants were handed a rating sheet, rating questions, and the Real test booklet. Questions to address for the Real test rounds were also handed out in survey format. In addition to the modified Angoff task and rank ordering of items in terms of difficulty, the participants were asked a couple of questions to test: (1) their capacity to recall replicated practice test items and, (2) recognition of similarity between the Practice and the Real test items in terms of knowledge and skills measured (refer to Appendix C for the questions the participants addressed in this first round of judgment of the Real test items). For instruction, the participants were told that the same judgment conventions as the Practice test applied to the Real test, they were asked to read instructions carefully before addressing questions, to ask for clarification if they get stuck, and to take down notes especially of their judged proportions because it would facilitate their table group discussion soon after. The participants were also asked to turn in their rating sheets when done and to hold on to the Real test booklet for the second round of judgment. After the first round of judgment on the Real test booklet, the participants took a short break. When they returned from the break, they were given guidelines with which to discuss their first round judgments in their table groups.

The seventh training activity was the table group discussion of the first round judgments of the Real test items. The guidelines for discussion were projected on the overhead as well as handed out to the participants. The participants were also given post it notes to reflect their thoughts during the discussions. The instructions on the discussion guidelines emphasized equal participation and sharing of insights, focus on judgments of the greater grey areas, discussion of constructs measured by the items, the modified Angoff strategy, sharing rationales, and noting changes to first round Real test recommendations. When the Heuristic training participants completed discussion of their first round recommendations, they were then handed the round two rating sheet. They were asked to adjust their judgments of proportions of barely proficient students that would respond correctly to items and item difficulty rankings if they wanted to, to reflect the information they acquired during discussion (refer to Appendix C for the discussion guidelines).

The eight training activity was the second round of judgment on the Real test. For this round, the participants were also handed rating sheets and rating questions. The participants were instructed to reflect on their earlier judgments and on the information gathered from the discussion and to make adjustments to their judgments if need be. The participants were given the opportunity to re-rank order items and re-judge the proportion of the participants that were barely proficient that would respond correctly to the Real test items. When they were done with this second round of judgments, evaluation forms were passed out for them to complete. Meanwhile, the participants were curious to see their first and second round of Real test judgments of the proportions of barely proficient students that would respond correctly to test items. As a consequence, both rounds were analyzed and presented to them after they completed the evaluation questionnaire.

The ninth and last training activity was filling out of the evaluation questionnaire. The evaluation questionnaire included questions to measure participant's satisfaction with the training, their engagement and emotion during the meeting, to ascertain the factors they considered in their judgment of the proportion of students that would get each item correctly, perception of confidence in judgments, and understanding of the training instructions. The participants were instructed to address all questions and to offer suggestions for future studies (please see Appendix C for the evaluation survey). The meeting concluded on a good note with thank you speeches, payment of stipend, and farewell messages.

6.4.2. Procedures for the Normative Training

The Normative training also began with introductions led by the researcher. Name tags and Power Point slides were used to facilitate these introductions. The facilitators introduced themselves by indicating their names, affiliation, and area of expertise. After the facilitators introduced themselves, the participants were asked to introduce themselves by stating their name, background, why they agreed to participate in the meeting, and what they expect to get out of the meeting. The latter two questions were asked also in order to gauge participant's motivation.

Following introductions, the researcher briefed participants about some housekeeping matters, an essential part of which was completing the Institutional Review Board (IRB) consent form. After going through house-keeping matters, the co-facilitator took over leading training discussions. The introductory part of the training entailed review of the purpose and agenda of the meeting, and these were projected on the overhead as well as read out from the script by the co-facilitator.

The first training activity was the review of background of the subset of the MEAP mathematics tests that we were used for the study. For review of the background of MEAP tests, the participants were given handouts delineating the content strands and GLCE's measured by the items on the Practice and the Real tests. The training instructions specified that the content area of the tests was mathematics, that the grade level of the tests was fourth grade, the items on the tests were multiple choice, there were 15 items on each of the tests selected from past MEAP test and that they were of varying difficulty levels, the content strands, and the GLCE's measured by the tests. The instruction also specified that the items were pilot tested and reviewed through a rigorous process so that their quality was not in doubt. These instructions were projected on the overhead and read out from the script (refer to Appendix B for the script read out to the Normative training participants for review of the background of the Practice and the Real tests). At the end of review of the test content material, the participants were encouraged to ask questions before proceeding with test administration.

The second training activity was taking the test. The participants were handed the Practice test booklet and answer sheet. Instructions on taking the test were projected on the overhead as well as read out from the script. The participants were told that it was essential to take the test to understand the experience that the fourth grade students have and to appreciate the content and demands placed on them. Besides, it was essential for them to become familiar with the tests for which they were to recommend cut scores. They were instructed on taking the test, to read each question carefully before responding, to think about how the fourth grade student would respond to each of the test items and to note comments they may have about the questions. When they completed taking the test, they turned in their answer sheets and took a short break. When the participants returned from the break, they were given the answer key and this completed the test taking session. Please note that it is not typical to collect answer sheets from participants. However participants were asked to do so for this dissertation study in order to gauge how they performed which is to also provide validity evidence on their eligibility to render the judgments.

The third training activity was the review of the Proficient performance level descriptors (PLD). For the PLD review, each participant was given a hard copy of the Proficient PLD, and flipcharts were handed out to each table group. The participants were instructed to elaborate on the PLD by giving three illustrations for each descriptor of what a barely proficient student should know and be able to. The instructions on how to elaborate on the Proficient PLD were projected and partly read out from the script (please note that the script was not completely followed for the PLD review). Following table group discussions, each group presented their work, during which the co-facilitator also led discussion to highlight similarities and differences in the table group descriptions of the barely proficient. The group discussion continued until it was gauged there was consensus among participants about the knowledge, skills, and abilities of the barely proficient students. Then, the participants were allowed to take a short break.

The fourth training activity was review of the modified Angoff instructions. For this review, the instructions were projected, handed out as well as read out from the script (please refer to Appendix B for the complete modified Angoff instruction script used for the Normative training). Meanwhile, the most important part of the modified Angoff method script read out to participants was as follows and italicized to distinguish it from the rest of the discussions:

The complete instructions to guide you with using this procedure to rate items are contained in the modified Angoff rating instructions hand out that I gave and are:

A. think about a classroom made up of 100 barely proficient students

For each item on the test:

B. Based on description of barely proficient students, what proportion of the students in the above classroom would answer the item correctly?

C. Mark the percentage from 0 to 100

Refer to the paper versions of these instructions and all materials provided in this training in making your judgments. Give your best informed judgments but do not agonize over them. Any questions?

The fifth training activity was the Practice test round of judgment. The practice round of judgment was based on the Practice test. Questions to address for the practice round of judgment were handed out to participants and in survey format along with rating sheets, and the Practice test booklets. The practice questions included rank ordering items in terms of difficulty and judging item difficulties for the barely proficient students (refer to Appendix C for the exact questions the Normative training participants addressed in the practice round). For instruction, the participants were asked to read directions carefully before addressing questions, to ask for clarification, to turn in their rating sheets when done, and to hold on to the Practice test booklet when they had completed the practice exercises for feedback. Please note that it is not typical to ask participants of modified Angoff studies to rank order items in terms of difficulty, however this exercise was incorporated to facilitate comparison of the Normative training outcome with those of the Heuristic training. When the participants completed the practice exercises, they took lunch break. While the participants were at lunch their Practice test judgments of the proportion of the barely proficient that would respond correctly to the items were analyzed. They were analyzed by the researcher and assisted by a fellow student using Excel and SPSS spreadsheet packages. Line charts were generated for each item for judgments of the proportion of the barely proficient that would respond correctly to items and for the cut score estimates of the participants. The cut score recommendation of the participants were estimated as the sum of their item level judgments of the proportion of the barely proficient that would respond correctly to the test item.

The sixth training activity was feedback on practice round of judgments. The participants were given feedback on their judgments of the proportion of the barely proficient that would respond correctly to each item and cut scores. From this practice feedback henceforth, the researcher took over facilitation of the study. The script was also followed for the feedback discussion. The feedback instructions emphasized that the goal was for the participants to compare their judgments and share rationales for recommendations and not to reach consensus. They were also instructed that the feedback would focus on line chart summary of their item difficulty judgments and cut scores. Line chart summaries of three items were displayed per slide, in order to make it easy for participants to digest the information. First, line chart summaries of participants item difficulty judgments were displayed that contained on the horizontal axis the participants' ID and on the vertical axis, their judgments of the proportion of barely proficient that would respond correctly to the items. The line chart displays also showed summary statistics of the item difficulty judgments (mean, standard deviation, minimum, maximum, and range of the judged proportions). Discussion of the line chart summaries of judged proportions of the barely proficient that would respond correctly to the items was process-oriented and structured to focus on items for which there was greater discrepancies of the judged proportions. Participants with outlying judgments were asked to share their rationales and the whole group deliberated on these views for a while before continuing to the next line chart display. This process was carried through for all such items and at the end of the deliberation the participants were told the empirical proportion of the fourth graders getting each of the items correctly in 2005. The second data summary displayed was the line chart of the cut score estimates of the participants for the Practice test. The line chart contained on the horizontal axis the participants' ID and on the vertical the cut score estimates along with the associated group summary statistics (mean, standard deviation, minimum, maximum, and range). The same process of asking participants with outlying estimates to share their rationales was followed for deliberation on the cut scores. Participants were curious to know how the cut scores were computed as a consequence the researcher also discussed the cut score computation process. At the conclusion of the feedback discussion, the participants were given a handout containing the rank ordering of items and the empirical proportions of the fourth grade students responding correctly to the practice items. After the practice round feedback, the Practice test booklets were collected from the participants and the standard setting activities based on the Real test booklet resumed.

The seventh training activity was the first round of standard setting judgments based on the Real test booklet. For this first round of judgment, the participants were handed rating sheet, rating questions, and the Real test booklet. Questions to address for the Real test rounds were also handed out in survey format. In addition to the modified Angoff task and rank ordering of items in terms of difficulty, the participants were asked a couple of questions to test: (1) their capacity to recall replicated practice test items and, (2) recognition of similarity between the Practice and the real test items in terms of knowledge and skills measured (refer to Appendix C for the questions the participants addressed for the Real test rounds of judgment). For instruction, the participants were told that the same judgment conventions as in practice round should be applied to the Real test, they were asked to read instructions well, ask for clarification, to note items of greater uncertainties, to take down their judged proportions on the Real test booklet for reference during construct map feedback, and to turn in their rating sheets when done. After the first round of the Real test judgment, the participants took a break during which their first round of judgments on the Real test was analyzed. Their first round cut score estimates were also computed as the sum of the item level judgments of the proportion of barely proficient that would respond correctly to the real test items. The first round Real test cut score estimates of the participants was projected on the over head with their identification number shown beside their estimate for the construct map feedback session.

The eight training activity was a reflection on the construct map feedback. When participants returned from the break, they were each handed a construct map. The construct map handout contained computed percentages of students responding correctly to the items on the Real test at some ability levels that corresponded with the Rasch model difficulty of items measuring the knowledge and skills of the Proficient performance level descriptor (refer to Appendix C for the construct map feedback). First, the information on the construct map was described then the participants were instructed to take note of their first round Real test cut score estimate, to look for the construct map cut score closest to theirs, to check how the rank order and absolute value of their item difficulty judgments correspond at that point, and to think about the possible reasons for these discrepancies. They were to reflect on these for a while and then may change their recommendations albeit they were not mandated to do so. During this reflection, the Normative training participants were allowed the opportunity to discuss with their table group members using the construct map feedback.

The ninth training activity was the second round of judgment on the Real test. For this round, the participants were also handed the rating sheets and rating questions. The participants were instructed to follow the same judgment conventions as in the practice and first round of the Real test judgment, to reflect on their round one Real test judgments and to consider adjustments to their round one judgment in the light of the information gleaned from the construct map feedback. The participants were allowed the opportunity to re-rank order items and re-judge the proportion of the participants that were barely proficient that would respond correctly to the Real test items. When they were done with this second round of judgments evaluation forms were passed out for them to complete.

The tenth and last training activity was filling out of the evaluation questionnaire. The evaluation questionnaire included questions to measure their satisfaction with the training, their engagement and emotion during the meeting, to ascertain the factors they considered in their judgment of the proportion of students that would get each item correctly, perception of confidence in judgments, and understanding of the training instructions. The meeting concluded on a good note with thank you speeches, payment of stipend, and fare well messages.

6.5. Empirical Data

The data analyzed for this dissertation study were as follows:

- The participant responses to information sheet questions on their demographics, experiences, and motivations
- The Likert scale evaluation survey questions on engagement, emotion, confidence, understanding, satisfaction with training and facilitators, and factors considered in

judgment of the proportion of the barely proficient students that would respond correctly to test items

- The open-ended evaluation survey questions on comments about adequacy, satisfaction with, and appropriateness of training procedures
- The Heuristic training designations of the Practice test items to content strands, GLCE, and DOK levels; and the Normative training Practice test response data
- The Heuristic and the Normative training responses to the questions testing their recall of the Practice test questions and recognition of similarity of the knowledge and skills measured by the Practice and the Real test questions
- The Practice test and the two rounds of Real test modified Angoff judgments. Precisely, the 10 participant × 15 items and the 12 participant × 15 items matrices of judgment of the percentages of students that would respond correctly to items for the Heuristic and Normative training, respectively
- The Practice and the Real test rounds of difficulty rank ordering of items for both studies. Precisely, the 10 participant × 15 items and the 12 participant × 15 items matrices of the difficulty rank ordering of items for the Heuristic and the Normative training, respectively

6.6. Empirical Data Analysis

The primary data to be evaluated are participant's absolute item difficulty estimates, rank ordering of items in terms of difficulty, and cut scores estimates. Item difficulties were estimated on the percentage scale (0-100) but were converted to decimals for analysis. The estimates were in response to the question of the percentage of fourth graders that are barely proficient that would respond correctly to the item. Cut scores were estimated as the sum of the proportion (0-1) scale for each participant and average for the group.

Data analysis conducted includes exploratory, descriptive, and inferential statistics. However, the main aim of analysis was to understand the data structure and in relation with the heuristic model assumptions. The Heuristic and the Normative training data were summarized using both principal coordinate analysis technique (PCOA) and indices of correlation for completeness of understanding. The PCOA analysis was used to investigate for each group and across rounds, if the data matrix of probability judgments can be summarized meaningfully in terms of relatively small number of dimensions so that the data can be visualized in a lower dimensional space. Participant's probability judgments and cut scores were also evaluated for reasonableness in relation to Rasch model calibrated item response bootstrap resampling cut score criterion measures and to evaluate the claim that the overall cut score reflects the Proficient performance PLD.

To summarize, there were three parts to data analysis and were as follows:

(1) Data reduction with multivariate statistics technique of Principal Coordinates Analysis (PCOA)

(2) Computation of correlation and cut score indexes to investigate relationship of outcomes of the modified Angoff studies with internal and external criteria. The internal criterion is the group item means of the judged proportion of the barely proficient that would respond correctly to the items while the external criteria were the empirical proportions of correct responses of the fourth grade students to the items in 2005, DOK designations of the items by content experts in 2005, and the Rasch model based PLD cut score estimate (see

description of the PLD cut score estimate in the immediately following section and also in the result section).

(3) Independent sample *t*-tests of significance of the difference in average correlation and cut score estimates of the Heuristic and the Normative training

6.7. The Hypothetical Proficient PLD Cut Score Estimate

The bootstrap approach to estimation of the PLD cut score was informed by principles of exemplar and prototype theories of categorization, sampling theory, and item response theory[8]. Threshold mastery of the fourth grade MEAP proficient knowledge and skills was operationalized as a 50:50 chance of correct response to the items measuring the PLDs. Released in 2005 fourth grade MEAP items were selected by the researcher that were judged to measure the PLDs. The hypothesized proficient PLD cut score on the Rasch model ability scale (θ scale) is the mean of 1000 means of bootstrap samples taken from the Rasch model estimated difficulties of the items that were judged to measure the PLD. The cut score on the test scale (a.k.a. true score) was estimated for the Practice and the Real test as the sum of the probabilities of correct response to test items with the probabilities of correct response estimated by plugging into the Rasch model, the bootstrap estimated proficient ability estimate for θ and Rasch difficulty estimates of the items based on fourth grade students responses to the items in 2005 for β.

[8] The reader could refer to Chapter two for categorization theories

Chapter Seven: Results

This Chapter presents results of data analysis conducted to address the research questions about the impact of the Heuristic training and of instruction and practice activities versus feedback on the substantive meaningful of judgment process and the technical quality of outcomes. The results presented also pertain to test of plausibility of hypotheses specified in Chapter four that delineated the conceptual framework of this dissertation.

There are two broad categories of evidence presented in this chapter and in logical order as follows: tests of balance of the Heuristic and the Normative training groups on potential confounding factors and evidence about the impact of the Heuristic training and of instruction and practice activities versus feedback. The results of comparisons of the Heuristic and the Normative training groups on the measured potential outcomes of the training interventions are presented in the logical order according to Kirkpatrick's (1994) ordered categories of measures for evaluating training. However, evidence in the Kirkpatrick's framework category of potential costs and benefits of the Heuristic training are excluded in this chapter, instead they are addressed in the discussion chapter along with the recommendations for future practice.

For the purpose of use of the Kirkpatrick's training evaluation framework for this dissertation, evidence in the categories of knowledge and skill acquisition and transfer of learning are presented together for convenience in one subsection and are broken further into specific types. However, it is important to highlight that the results presented in the category of performance on pre-requisite tasks of categorization, recall, and rank ordering of items in terms of difficulties is the knowledge and skill acquisition evidence while indicators of performance on the modified Angoff task are the transfer of learning evidence. It is important to highlight also that in practical implementations of the Heuristic training, follow up evaluation evidence can be sought for as to whether what the participants of the Heuristic training learned transferred to their classroom practice. Also, in presenting the evidence about performance on the modified Angoff item difficulty judgment task, reference is made to the Kane's (2001) and Raymond and Reid's (2001) types of evidence which also fall into Kirkpatrick's transfer of learning category when they come up.

7.1. Potential Confounds

The conceptual framework of this dissertation presented in Chapter four identified categories of extraneous factors that could also impact knowledge and skills acquisition and the probability judgment outcomes apart from the Heuristic training interventions[9]. Specifically, three categories of extraneous factors were identified in the conceptual framework and were:

- Non-cognitive constructs attributes of participants
- Background characteristics of participants
- Training procedural implementation factors

The first analytic task was to check for balance of the Heuristic and the Normative training group on the measured instances of the aforementioned categories of potential confounding factors. In the category of training procedural implementation, the facilitators and stimulus materials for both studies were the same so that they are ruled out as potential confounds. The potential confounding of training procedural implementation factors such as order effects are addressed in Chapter eight in the presentation of the limitation of this dissertation. Meanwhile the results of independent sample *t*-tests conducted to check for balance

[9] Extraneous variables are variables that are not of interest to the researcher but which could potentially confound with the outcome of the training intervention

of the Heuristic and the Normative training groups on the measured non-cognitive constructs and background characteristics are presented in Table 7-1.

Table 7-1 presents the summary statistics of the measured extraneous variables and as identified by the conceptual framework of this dissertation that could potentially confound with the impact of the Heuristic training on judgment outcomes. The column labeled "Study group" specifies the training groups, "Variable" indicates the variable for which statistics are presented, "*N"* specifies the sample sizes of the study groups used for computing the statistics for the involved variable. The sample size of the Normative training group is (*N=12*) and the sample size of the Heuristic training group is (*N=10*). If the sample size is less than the training group size for a given variable, it implies that the variable had missing values. For instance, the number of years of experience teaching had considerable number of missing values because the researcher missed out asking current teachers to indicate how many years they had been teaching[10]. The columns with symbol "*M*" and "*SD*" stands for the means and the standard deviations of the variables, respectively.

In the category of background characteristics, the indicator variables for which statistics are presented in Table 7-1 include: gender (Female), teaching experience (Taught), teaching experience at third or fourth grade (Taught third or fourth grade), experience teaching math (Taught math), and math specialty (Math specialist). For the indicator variables the sample sizes for the "Yes" and "No" categories are also specified in the column labeled "*N*" that contains the sample size of the training studies used for computing the statistics for the involved variable. Also presented in Table 7-1 are the statistics for background characteristics measured on the

[10] This also means that the values are not missing at random so that the measure is somewhat biased, however is included for completeness of discussion

interval scale namely: number of years in educational field (No. yrs. in educ. field) and number of years teaching (No. yrs. teaching)[11]. In the category of non-cognitive constructs, the variables for which statistics are presented in Table 7-1 include: intrinsic motivation, extrinsic motivation, emotion, and engagement.

The results of independent samples *t*-test check for balance of the Heuristic and the Normative training groups on background characteristics and non-cognitive constructs is presented in the Table 7-2. The column labeled “Variable” indicates the variable for which independent samples *t*-tests are presented, “Estimate” refers to the mean difference between the Heuristic and the Normative group on the involved variable, “S.E.” the standard error of the mean difference, “*t*-value” is the *t*-statistic, and “**P(**$|t|$ **>*t*)**” is the p-value. Except for the variables in which there were missing values, the *t*-distribution is $|t|(20)$[12]. The equal variance assumption of both groups on the measured variables was met. The mean difference between the Heuristic and the Normative training groups on all of the variables were statistically insignificant at an α *=.1*.

[11] In parenthesis are the labels given to the variable in Table 7-1

[12] The degree of freedom for the *t*-tests in general is the sum of the sample sizes minus 2

Table 7-1: Distribution of Measured Extraneous Variables

Category	Study	Variable	*N*	*M*	*SD*
Background Characteristics	Heuristic	Female	Yes = 7	.70	.48
			No = 3		
	Normative	Female	Yes = 10	.83	.39
			No = 2		
	Heuristic	Taught	Yes = 8	.80	.42
			No = 2		
	Normative	Taught	Yes = 9	.75	.45
			No = 3		
	Heuristic	Taught third or fourth grade	Yes = 5	.50	.53
			No = 5		
	Normative	Taught third or fourth grade	Yes = 3	.25	.45
			No = 9		
	Heuristic	Taught math	Yes = 8	.80	.42
			No = 2		
	Normative	Taught math	Yes = 7	.58	.52
			No = 5		
	Heuristic	Math specialist	Yes = 6	.60	.52
			No = 4		
	Normative	Math specialist	Yes = 8	.73	.47
			No = 3		
	Heuristic	No. yrs. in educ. Field	10	11.50	8.61
	Normative	No. yrs. in educ. Field	12	7.17	9.05
	Heuristic	No. yrs. exp. Teaching	7	7.14	8.93
	Normative	No. yrs. exp. Teaching	9	3.00	3.39
Non-Cognitive Constructs	Heuristic	Extrinsic	9	9.67	2.18
	Normative	Extrinsic	12	9.08	1.56
	Heuristic	Intrinsic	10	12	2.16
	Normative	Intrinsic	11	11.73	2.83
	Heuristic	Emotion	9	10.44	1.51
	Normative	Emotion	12	10.33	1.23
	Heuristic	Engagement	9	17.89	1.69
	Normative	Engagement	12	18.25	1.71

Table 7-2: Independent Sample *t*-Tests for Extraneous Variables

Category	**Variable**	**Estimate**	**S.E.**	***t*-value**	**P($\|t\|$ >*t***
Background Characteristics	Female	-.13	.19	-.72	.48
	Taught	.05	.19	.27	.79
	Taught third or fourth grade	.25	.21	1.20	.25
	Taught math	.22	.20	1.07	.30
	Math specialist	-.13	.22	-.59	.56
	No. yrs. in educ. Field	4.33	3.79	1.14	.27
	No. yrs. exp. teaching	4.14	3.22	1.29	.22
Non-Cognitive Constructs					
	Extrinsic motivation	.58	.82	.72	.48
	Intrinsic motivation	.27	1.11	.25	.81
	Emotion	.11	.60	.19	.85
	Engagement	-.36	.75	-.48	.64

7.2. Participants Reaction

The hypothesis for which the results are reported in this section is that based on Kirkpatrick's category of participant's satisfaction. This level of framework measure suggests that a qualitatively better training should result in more positive reaction from the participants. The results indicated that the participants of the Heuristic training expressed higher satisfaction on the measures that elicited their reaction than those in the Normative training. The evidence on the basis of which this conclusion was based are percentage scale summaries of their Likert scale questions responses, independent sample *t*-tests of mean differences, and open ended item responses of the participants. This evidence is also part of Kane's (2001) procedural validity evidence. Results are presented in this section beginning with the percentage scale summary of the responses of the participants to Likert scale items with associated independent sample *t*-tests, and then followed by their responses to open ended items.

Table 7-3 presents summary in percentage points of the Heuristic and the Normative training responses to the Likert scale question that asked about their overall assessment of the training. As shown in Table 7-3, (70%) of the Heuristic group versus (33%) of the Normative training group expressed that the training was very good. Independent samples *t*-test was used to examine the mean difference between the Heuristic and the Normative training groups on the extent to which they were satisfied with the training. The Levene's test of equal variance showed insignificant difference of the variation in the expressed satisfaction of the Heuristic and the Normative groups with the training intervention *F (1, 20) = .44, p=.52.* The Heuristic training group expressed a higher mean satisfaction with the training (*M = 3.70, SD = .48*) than the Normative training group (*M = 3.25, SD = .62*). The mean difference between the two groups expressed satisfaction with the training intervention was statistically significant at α *=.1,* $|t|$ *(20) =1.87, p = .08.*

Table 7-3: Participants Assessment of Training

	Poor	**Fair**	**Good**	**Very Good**
Heuristic	0	0	30	70
Normative	0	8.3	58.3	33.3

Table 7-4 presents summary in percentage points of the Heuristic and the Normative training group's responses to the question that asked the participants' overall assessment of the facilitators of the training. As shown in Table 7-4, (80%) of the Heuristic group versus (18%) of the Normative group participants graded the facilitation as very good. Independent sample *t*-test was used to examine the mean difference between the Heuristic and the Normative training groups on the extent to which they reacted positively to the training facilitation. The Levene's test of equal variance showed insignificant difference of the variation in expressed satisfaction of the Heuristic and the Normative training groups with the facilitators *F(1,19) = .005, p=.95.* The Heuristic training group expressed a higher mean (*M = 3.80, SD=.42*) satisfaction with the facilitators than the Normative training group (*M = 3.09, SD=.54*). The mean difference between the two groups was statistically significant at α *=.1,* $|t|$ *(19) = (3.33), p = .004*.

Table 7-4: Participants Assessment of Facilitators

	Poor	**Fair**	**Good**	**Very Good**
Heuristic	0	0	20	80
Normative	0	9.1	72.7	18.2

The rest of the discussion in this section pertains to summary of the responses of the participants to the open-ended items that elicited the reactions of the participants to different aspects of the training. The responses of the participants to the open-ended question of what they found helpful about the training are presented in Tables 7-5 and 7-6, for the Heuristic and the Normative training groups, respectively. In Tables 7-5 and 7-6, short responses were reported verbatim while long responses were summarized to highlight on-message comments. The summary responses indicated that participants valued most the discussions and interactions that took place during the training. Their responses also suggested helpfulness of the Angoff method.

Table 7-5 presents summary responses of the Heuristic training group to the open-ended item that asked them to indicate what they found most helpful about the workshop. Overall the Heuristic training groups' responses to this question showed positive perceptions of the training. Except for one participant who expressed that the training did not give insight about the standard setting process used by the Michigan Educational Assessment Program (MEAP), all commented positively about the training. For the most part, participant's responses to this open-ended item fell into two categories namely: training discussions and the modified Angoff process, and are presented in Table 7-5.

Table 7-6 presents summary responses of the Normative training participants' responses to the open-ended item that asked them to indicate what they found most helpful about the workshop. Overall the Normative training group's responses to this question also showed positive perception of the training. All had at least a positive comment about the training. The Normative training participants responses also mostly fell into the same two categories as those of the Heuristic group namely: training discussions and the modified Angoff process, and are presented in Table 7-6.

Table 7-5: Heuristic: What Did You Find Most Helpful About the Workshop?

Category	**Example Responses**
Discussions	The most helpful part of the training was the small group breakouts. These provided an opportunity to share among a small group of individuals and then compare among the other groups. The practice section was critical because it allowed the opportunity to clarify questions that I had. The practice and discussions that followed It was very interesting to go through the process for determining cut scores. It was also very helpful to have the PLD discussion to get a better understanding of what it takes to be proficient on a particular objective I really valued the discussions we had in small and large groups. They helped to norm my responses as to what proficiency really looks like
The Angoff Process	I found looking at questions and developing cut scores to be an interesting process. How to arbitrarily determine "Proficient to Not" really takes a lot of thought and careful analysis The process to found the cut off score. Thinking about each question how many barely proficient students would get a correct? How other teachers think about barely proficient students and their thinking In general, the process of seeing what goes in to analyzing a cut score was very helpful. Seeing multiple rounds of scoring was useful as well.

Table 7-6: Normative: What Did You Find Most Helpful About the Workshop?

Category	Example Responses
Discussions	I liked being able to discuss and analyze data with colleagues. I found the whole group discussions the most helpful. Talking to other educators, especially those closer to the appropriate level, helped to clear up some of my thoughts and concerns I found the table discussions very helpful to see collectively how students may answer a question Being able to work with people in different levels of the field. This helped me understand better and what is expected of a student to find the cut score I enjoyed the opportunities to talk to other people and discuss a subject area they were passionate about. I enjoyed the discussions that took place in the very beginning of the workshop which covered the topic of what is considered "very-proficient". I found this most interesting because it is challenging to determine a cut-off since you strive for something higher than "barely-proficient" I found the PLD discussions, table discussions, whole-group discussions and the exercises extremely helpful The Group discussions were the most interesting, helpful parts about the training
The Angoff Process	I learned how cut scores are attained (one way) and feel the strategy (Angoff) is something I could possibly use as assessment of my instruction/need to teach The exercises actually doing the rating/then comparing results and discussing Experiencing how cut scores are determined I have never experienced anything like this, taking an in-depth look at assessments and realizing how much work goes into creating them. It was very interesting I really enjoyed looking at various questions and how in depth their complexity is. It was also a change of pace to look into elementary testing and not just for high school

Tables 7-7 and 7-8 shows summary of responses of the participants to the question on what they would have liked to know more about in the training. It can be observed from Tables 7-7 and 7-8 that the participants of both the Heuristic and the Normative training expressed that they would have liked to know more about other standard setting methods for determining cut scores, especially the method applied by the Michigan state government for the MEAP test. In

Tables 7-7 and 7-8, the same convention of reporting short responses verbatim and on-message aspects of long responses was followed.

Table 7-7 shows summary of the responses of the Heuristic training group to the question that asked them to indicate what they would have liked to know more about in the training. With the exception of one participant who expressed dissatisfaction with the explanation of graphical displays of feedback data summary, the responses to this question fell into two categories namely: the cut score estimation methods and MEAP method, and the test development process and assessment categories.

Table 7-8 presents responses of the Normative training group to the question that asked them to indicate what they would have liked to know more about in the training. Except for a couple of remarks about the lack of clarity of instructions on how to elaborate on the performance level descriptors and on the construct map feedback, the Normative training participants responses also fell mostly into the same two categories namely: the cut score estimation methods and MEAP method, the test development process, and assessment categories.

Table 7-7: Heuristic: During the Training What Would You Have Liked to Know More About?

Category	Example Response
Cut Score Estimation Methods and MEAP Method	Cut Scores and how the state uses them I would have liked to learn more about other processes for setting cut scores. This is a fascinating process that I did not know a lot about ahead of time To sum the scores of each item to get the cut score is one strategy. Are people also think about using normal distribution method to work on this? What are other strategies that have been applied to practice? E.g. How the new cut scores of MEAP are calculated? Other methods of determining cut scores. How the state chooses the "experts" for the real life scenarios of determining cut scores for MEAP What criteria do the actual MEAP testers use to determine their cut scores? (prior data? Demographics? Etc.
Test Development Process and Assessment Categories	Writing the actual assessment questions to target the learning objective How will the common core affect the test questions? Will questions be deeper in 3-4 topics? Less questions, but more story problems? Comparing MEAPs since 2005-have they gotten harder, easier, or about the same? Ask a MEAP test writer or evaluator to speak about process in writing questions and the language used. What makes a good MEAP question? I could of used a better introduction or explanation of the DOK sheet because I was unaware of it

Table 7-8: Normative: During the Training, What Would You Have Liked to Know More About?

Category	Example Response
Cut score Estimation Methods and MEAP Method	The behind the scenes mathematical analysis process would be fun to know. Also, just a brief description of other methods used to find cut scores I would have liked to know more about the construct map. I felt confused when it was given to me and I feel it would have helped to discuss our rankings as a group before given the students results I would like to know more about other methods of determining cut scores - just an overview The process that MDE actually uses I would liked to learn more about how the actual cut scores for different types of formal assessments are constructed (i.e. MEAP; ACT, etc.) The way cut score values are actually calculated and how we can change our assessments to better fit those projected cut scores The process of how the State of Michigan predicts cut scores
Test Development Process and Assessment Categories	I think I would like to know more about how the wording on the test could have influenced student answers It would have been helpful to not only know the grade 3 GLCE's but to know how long students have been studying each topic (when was it first introduced?).

In general, participants' responses to the question that asked whether they had questions or concerns that weren't addressed in the training, as presented in tables 7-9 and 7-10, for both groups, suggested that they were satisfied with the training and that their questions were adequately addressed. On the other hand, the researcher concurs with some of the participant's comments about the insufficiency of background information given about the study. It is noteworthy that these details were purposely left out in other to make the training as concise as possible.

Table 7-9 presents responses of the Heuristic training participants to the question that asked them to indicate if they had questions that were not addressed in the training. Six out of the 10 Heuristic training participants indicated that all their questions were adequately addressed. On the other hand, the issues raised by the remaining four fell mostly into two categories namely: background of the study (e.g. research questions and how the data will be analyzed) and the clarity of the methods and these are presented in Table 7-9.

Table 7-10 presents responses of the Normative training participants to the question that asked them to indicate if they had questions that were not addressed in the training. Seven of the 12 Normative training participants indicated that all their questions were adequately addressed. One of the 12 Normative training participants did not respond to this question. The rest of the Normative training participants raised issues that also fell into the same two categories as with the Heuristic training participants namely: background of the study (e.g. research questions and how data will be analyzed) and the clarity of the methods.

Table 7-9: Heuristic: Did You Have Questions or Concerns That Were Not Answered or Addressed in the Training Session?

Category	Example Response
Background and Purpose of Study	What's your research question? What made you choose this topic to investigate? How are out data regarding the GLCE's and DOK's going to be used? Why did we fill these out?
Clarity of Methods and Instruction	Where is policy heading with test design and how does what we did today reflect or interact with current trends in the field? A model of what we were supposed to be doing on those flip chart pages for the initial task would have been helpful. How were the original and new cut scores produced? I think I would like to be exposed to more real examples.

Table 7-10: Normative: Did You Have Questions or Concerns That Were Not Answered or Addressed in the Training Session?

Category	Example Response
Background and Purpose of Study	I was also slightly unclear on what specific topic this dissertation was going to address
Clarity of Methods and Instruction	My main concern was not understanding the construct map but talking afterwards helped to clarify Just the process that MDE uses and how it is different than what we did As mentioned above, I would have liked to learn more about cut scores. If the construct data table included students of all proficiencies, why were we able to change our scores?

Table 7-11 and Table 7-12 presents responses of the Heuristic training and the Normative training group, respectively to the question that asked them to provide additional comments concerning the adequacy, appropriateness, usefulness, or organization of the training. The Normative training group raised more issues than the Heuristic training group as can be observed from Table 7-12 which is a lot more elongated however, most of both training participants found the training useful and well organized.

Table 7-11: Heuristic: Please Use the Space Below to Provide Additional Comments Concerning the Adequacy, Appropriateness, Usefulness, or Organization of the Training.

Category	Example
Issues	I feel this is a very important area for discussion. As we saw, the differences in how we rated questions was wide but after training they were closer I would like to see this type of workshop using the new common core. Then it would be something we can go back and have in our hands as relevant I feel that we could have used more time in the beginning to discuss the PLDs. It was a valuable part of the day and felt rushed
Commendations	The training was well organized and very useful. I am extremely glad that I chose to participate! I actually quite satisfied with the arrangement It was very well done. Great training. Made me think more about how to get those potential barely proficient students to be proficient! I'll think about other ways to keep the students immersed in wanting to know more and was to remember concepts The flow of the training was evidently very organized. The training was concise, to the point and very educational Very well organized; some directions were unclear at the beginning of each session, but the overall structure and flow of the day was good. Scheduled more than enough time to complete the session

Table 7-12: Normative: Please Use the Space Below to Provide Additional Comments Concerning the Adequacy, Appropriateness, Usefulness, or Organization of the Training

Category	Example
Issues	As the MEAP does, doing a group sample question before beginning each task would have been very helpful for us visual/hands on learners It would have been great to have an idea of how to use this in classroom instruction, from a teacher stand point Directions could have been clearer at times I thought the set up and agenda helped me through the day. I wish we didn't have as many papers as we were given and more so a packet all at once and that way use less Tasks were interesting. Sometimes directions weren't clear as to what we would be doing. It was very timely and kept moving It was hard to understand what was expected of us in the discussion of PLDS I think your timing of activities was off and could be adjusted I think it was useful for me as a teacher to start to think about having different expectations for different students based on their abilities and potential I enjoyed this activity but didn't necessarily think it pertained to my teaching background or high school students
Commendations	This was a good training to get a snapshot of the process and a better understanding. Very organized. Very methodical. Facilitators able to instruct and answer all questions Very organized. The training program was very well-organized and I felt that a lot of work was put into preparation. You have inspired me to do some research of my own into the specific of cut scores The training was fun. I felt everything was very well organized.

7.3. Knowledge and Skill Acquisition and Transfer to Tasks

For the rest of the discussion of the results, the measurement terminology item difficulty will be used instead of probability judgments. Item difficulties will be used in order to make the relationship of the task of judging conditional probabilities of correct response for test items and those of the intermediate tasks of categorizing, and rank ordering items in terms of difficulty transparent.

This section presents evidence addressing the fundamental research questions and hypotheses of this dissertation. Specifically, the claims that the Heuristic training would result in higher conceptual understanding and consideration of the knowledge and skills constructs measured by the tests in the item difficulty judgment process, yield judgments that are more substantively meaningful, and of higher technical quality. The analyses results reported are descriptive statistics and independent sample *t*-tests of significance of difference in the average performance of the Heuristic and the Normative training participants on the pre-requisite tasks, and on the modified Angoff item difficulty judgment tasks.

There are five sub-sections in this part, each presenting a different type of evidence about participant's knowledge and skills and transfer of knowledge and skills to the modified Angoff item difficulty judgment tasks. The self-report of the participants about their task performance is presented first, followed by indices of their performance on the posited pre-requisite knowledge and skills tasks, and then indices of their performance on the modified Angoff item difficulty judgment task. Specifically, the five types of evidence presented in the subsections that follow and in order are: (1) self-reports of task performance; (2) performance on item difficulty judgment pre-requisite tasks; (3) exploratory analysis of judged item difficulties data; (4) quantitative indices of relationship of judged item difficulties with internal and external criteria; and, (5) quantitative indices of derived cut scores location, variability, and stability across rounds of judgment. All performance indicators pertaining to the practice round of item difficulty judgment provides direct evidence about the impact of the Heuristic training instructions and practice activities[13]. However, comparative performance of the Heuristic and the Normative training participants in the feedback rounds of judgment also provide evidence albeit indirect,

[13] i.e., effect of training instructions prior to and including the practice round of judgment

about the progressive impacts of the differential instructional and practice activities for the two groups.

7.3.1. Self-Reports

The self-reports of tasks performance addresses the hypotheses that the Heuristic training would result in higher awareness of and consideration of knowledge and skills constructs measured by the tests in judgments, better understanding, confidence, and perceived competence in task performance. The evidence presented in this section includes percentage point summaries of the participant's responses to Likert scale items that elicited their perspective of how they performed on the tasks. Also, independent sample *t*-tests of the mean difference between the Heuristic and the Normative training participants' expressed weight given to experiential and knowledge and skills constructs in their judgment, perceived understanding of instruction and tasks, and confidence in recommendations are presented.

Table 7-13 presents summary in percentage points of the responses of the Heuristic and the Normative training groups to the four-point Likert scale item that asked them to indicate the extent to which discussion and feedback impacted their recommendations. As shown in Table 7-13, (75%) of the Normative training participants versus (40%) of the Heuristic training participants strongly agreed that discussion and feedback impacted their recommendations. Independent samples *t*-test was used to examine the mean difference between the Heuristic and the Normative training group participants expressed extent to which discussion and feedback impacted their recommendation. The Normative training participants expressed a higher mean (*M = 3.75, SD= .45*) impact of discussion and feedback on their recommendations than the Heuristic training group (*M = 3.30, SD=.68*). The Levene's test of equal variance showed insignificant difference of the variation in the responses of the two groups on the impact of feedback and discussion in their judgment deliberation *F (1, 20) = 2.45, p=.13.* The mean difference between the two groups was statistically significant at α *=.1,* $|t|$ *(20) =1.87, p = .08.*

Table 7-13: Discussion and Feedback Impacted Recommendations?

	Strongly Disagree	**Disagree**	**Agree**	**Strongly Agree**
Heuristic	0	10	50	40
Normative	0	0	25	75

Table 7-14 presents summary in percentage points of the responses of the Heuristic and the Normative training groups to the four-point Likert scale item that asked them to indicate the extent to which their educational or classroom experience impacted their recommendations. As shown in Table 7-14, (80%) of the Heuristic training participants versus nearly (42%) of the Normative participants strongly agreed that their educational or classroom experience impacted their recommendations. Independent samples *t*-test was used to examine the mean difference between the Heuristic and the Normative training groups on the extent to which their educational or classroom experience impacted their recommendation. The Heuristic training group expressed a higher mean (*M = 3.80, SD= .42*) impact of their educational or classroom experience on their recommendations than the Normative training group (*M = 3.17, SD=.84*). The Levene's test of equal variance showed significantly higher variation of the responses by the Normative training participants on the impact of their educational and classroom experience in their judgment deliberation than the Heuristic training group *F(1,20) = 6.23, p=.02.* The Heuristic training participants expressed statistically significant higher impact of their educational and classroom experience in their judgment deliberation than the Normative training participants, at α *=.1,* $|t|$ *(16.84) =2.30, p = .04.*

Table 7-14: Educational or Classroom Experience Impacted Recommendations?

Factors	Strongly Disagree	Disagree	Agree	Strongly Agree
Heuristic	0	0	20	80
Normative	0	25	33.33	41.67

Table 7-15 presents summary in percentage points of the responses of the Heuristic and the Normative training groups to the four-point Likert scale item that asked them to indicate the extent to which the knowledge and skills construct measured by the items impacted their recommendations. As shown in Table 7-15, (30%) of the Heuristic training participants versus (8%) of the Normative training participants strongly agreed that the constructs measured by the items impacted their recommendation. Independent samples *t*-test was used to examine the mean difference between the Heuristic and the Normative training groups expressed extent to which the knowledge and skills constructs measured by test items impacted their recommendation. The Heuristic training group expressed a higher mean (*M =3.30, SD = .48*) impact of the knowledge and skills constructs measured by the test items on their recommendations than the Normative training group (*M = 3.00, SD = .43*). The Levene's test of equal variance showed significant difference of the variation of the two groups expressed impact of the knowledge and skills constructs measured by the tests on their judgment deliberations *F(1,20) = 3.50, p=.08.* The mean difference in the expressed impact of the knowledge and skills constructs measured by the test items on their judgment deliberation by the two groups was statistically insignificant at α $=.1$, $|t|\,(18.20) =1.53, p = .14$.

Table 7-15: What is Measured by Items Impacted Recommendations?

Factors	Strongly Disagree	Disagree	Agree	Strongly Agree
Heuristic	0	0	70	30
Normative	0	8.3	83.33	8.3

Table 7-16 presents summary in percentage points of the responses of the Heuristic and the Normative training groups to the four-point Likert scale item that asked them to indicate the extent to which the quality of the items impacted their recommendations. As shown in Table 7-16, (50%) of the Heuristic training group versus (33%) of the Normative training group strongly agreed that the quality of the items impacted their recommendation. Independent samples *t*-test was used to examine the mean difference between the Heuristic and the Normative training groups on the extent to which the quality of test items impacted their recommendation. The Heuristic training group expressed a higher mean (*M =3.40, SD = .70*) impact of the quality of test items on their recommendations than the Normative training group (*M = 3.08, SD = .90*). The Levene's test of equal variance showed insignificant difference of the variation in the expressed impact of item quality on their judgment deliberation by the two groups *F (1, 20) =.003, p=.96.* The mean difference between the two groups was statistically insignificant at α *=.1,* $|t|$ *(20) =.91, p = .38.*

Table 7-16: Item Quality Impacted Recommendations?

	Strongly Disagree	Disagree	Agree	Strongly Agree
Heuristic	0	10	40	50
Normative	8.3	8.3	50	33.33

Table 7-17 presents summary in percentage points of the responses of the Heuristic and the Normative training groups to the four-point Likert scale item that asked them to indicate the extent to which the PLD impacted their recommendations. Independent samples *t*-test was used to examine the mean difference between the Heuristic and the Normative training groups on the extent to which the PLD impacted their recommendation. The Heuristic and the Normative training participants on the average expressed identical average impact of the PLD on their recommendation. In terms of specifics, the Heuristic response statistics were (*M = 3.00, SD = .47)* and the Normative responses statistics were *(M = 3.00, SD = .60)*. The Levene's test of equal variance showed insignificant difference of the variation of the expressed impact of the PLD on their judgment deliberation by the Heuristic and the Normative training groups *F(1,20) = .46, p=.51*. Likewise, the mean difference between the two groups perception of influence of the PLD on their judgment was statistically insignificant at $\alpha = .1$, $|t|(20) = .00, p = 1$.

Table 7-17: PLD Impacted Recommendations?

Factors	**Strongly Disagree**	**Disagree**	**Agree**	**Strongly Agree**
Heuristic	0	10	80	10
Normative	0	16.67	66.67	16.67

Table 7-18 presents summary result of the composite, sum score of the Heuristic training and the Normative training participants on all aforementioned five factors. The summary statistics in Table 7-18 are on the scale of 0-20. Independent samples *t*-test was used to examine the mean difference between the Heuristic and the Normative training groups on the extent to which all five factors as presented in Tables 7-13 through 7-17 impacted their recommendation. As shown in Table 7-18, the Heuristic training group expressed a higher mean (*M =16.80*) impact of all five factors on their recommendations than the Normative training group (*M =16*). The Levene's test of equal variance showed insignificant difference of the variation in the overall expressed impact of all five factors in their judgment deliberation by the two groups *F(1,20) =*

2.38, p=.14. The mean difference between the two groups was statistically insignificant at $\alpha = .1$, $|t|$ *(20) =1.30, p = .21.*

Table 7-18: Composite of Factors Impacting Recommendation

	Mean	**Range**	**Minimum**	**Maximum**	**Standard Deviation**
Heuristic	16.80	4	15	19	1.62
Normative	16	4	14	18	1.28

Tables 7-19 and 7-20 show summary statistics in percentages of the responses of the Heuristic and the Normative training participants to the questions aimed at eliciting their understanding of the tasks. Table 7-19 presents summary in percentage points of the Heuristic and the Normative training groups responses to the Likert scale question that asked if they were able to follow instructions and complete ratings accurately. As shown in Table 7-19, (60%) of the Heuristic training group versus (25%) of the Normative training group strongly agreed that they were able to follow instructions and complete ratings accurately. Independent samples *t*-test was used to examine the mean difference between the Heuristic and the Normative training groups on the extent to which they were able to follow instructions and complete ratings accurately. The Heuristic training group expressed a higher mean of been able to follow instructions and complete tasks accurately (*M = 3.60, SD=.52*) than the Normative training group (*M = 3.25, SD=.45*). The Levene's test of equal variance showed insignificant difference of the variation in expressed perceptions of the two groups on their ability to follow instructions and complete ratings accurately *F(1,20) =1.83, p=.19.* The mean difference in the expressed capacity to follow instructions and complete ratings accurately of the two groups was statistically insignificant although nearing significance at $\alpha = .1$, $|t|$ *(20) = 1.70, p = .11.*

Table 7-19: Could Follow Instructions and Complete Ratings Accurately?

	Strongly Disagree	Disagree	Agree	Strongly Agree
Heuristic	0	0	40	60
Normative	0	0	75	25

Table 7-20 presents summary in percentage points of the Heuristic and the Normative training group responses to the Likert scale question that asked if they understood tasks and feedback. As shown in Table 7-20, (50%) of the Heuristic group versus (33%) of the Normative group strongly agreed that they understood tasks and feedback. Independent samples *t*-test was used to examine the mean difference between the Heuristic and the Normative group on the extent of their perception of understanding of the tasks and feedback. The Heuristic training group expressed a higher mean perception of understanding of tasks and feedback (*M = 3.50, SD=.53*) than the Normative training group (*M = 3.25, SD=.62*). The Levene's test of equal variance showed insignificant difference of the variation of the Heuristic and the Normative training participants expressed understanding of tasks and feedback *F(1,20) = .00, p=1*. The mean difference between the two groups was statistically insignificant at α *=.1,* $|t|$ *(20) =1.01, p = .33*.

Table 7-20: Understood Tasks and Feedback?

Understanding	Strongly Disagree	Disagree	Agree	Strongly Agree
Heuristic	0	0	50	50
Normative	0	8.3	58.33	33.33

Table 7-21 presents summary statistics of the composite, sum of the scores of the Heuristic and the Normative training participants on the questions in Tables 7-19 and 7-20, which elicited their understanding of training tasks and feedback. The results presented in Table

7-21 is on the scale of 0 to 8. Independent samples *t*-test was used to examine the mean difference between the Heuristic and the Normative training groups on the composite of the questions in Tables 7-19 and 7-20 that elicited their understanding of tasks and feedback. The Heuristic training group expressed a higher mean perception of understanding of tasks and feedback (*M = 7.10, SD=.88*) than the Normative training group (*M = 6.50, SD=.80*). The Levene's test of equal variance showed insignificant difference in the variation of the Heuristic and the Normative training participants expressed understanding *F (1, 20) = .09, p=.77*. The mean difference between the two groups was statistically insignificant at α *=.1,* $|t|$ *(20) =1.68, p = .11*.

Table 7-21: Composite of Understanding of Task

	Mean	**Range**	**Minimum**	**Maximum**	**Standard Deviation**
Heuristic	7.10	2	6	8	.88
Normative	6.50	2	6	8	.80

Tables 7-22 and 7-23 show summary statistics of questions that elicited the Heuristic and the Normative participants' confidence in their recommendations. The breakdown of the responses of the participants to the confidence questions are summarized in percentages in Tables 7-22 and 7-23 that follow.

In Table 7-22, (60%) of the Heuristic training group versus (25%) of the Normative training group strongly agreed that they were confident about their conception of the barely proficient student. Independent samples *t*-test was used to examine the mean difference between the Heuristic and the Normative training participants on the extent to which they were confident about their conception of the barely proficient student. The Heuristic training group expressed a higher mean confidence in their conception of the barely proficient student (*M = 3.50, SD=.71*) than the Normative training group (*M = 3.17, SD=.58*). The Levene's test of equal variance showed insignificant difference of the variation of the Heuristic and the Normative training participants expressed confidence in their conception of the barely proficient student *F (1, 20) = 1.48, p=.24*. The mean difference between the two groups was statistically insignificant at α *=.1*, $|t|$ *(20) =1.22, p = .24.*

Table 7-22: Confident in Conception of the Barely Proficient Student?

	Strongly Disagree	**Disagree**	**Agree**	**Strongly Agree**
Heuristic	0	10	30	60
Normative	0	8.3	66.67	25

In Table 7-23, (40%) of the Heuristic training group versus (25%) of the Normative training group strongly agreed that they were confident in their cut score recommendation. Independent samples *t*-test was used to examine the mean difference between the Heuristic and the Normative training group on the extent to which they were confident about their cut score recommendation. The Heuristic training group expressed a higher mean confidence in their cut score recommendation (*M = 3.40, SD=.52*) than the Normative training group (*M = 3.00, SD=.74*). The Levene's test of equal variance showed insignificant difference of the variation of the Heuristic and the Normative training participants expressed confidence in their cut score *F (1, 20) = .01, p=.91*. The mean difference between the two groups was statistically insignificant at α *=.1,* $|t|$ *(20) =1.44, p = .17.*

Table 7-23: Confident in Cut Score?

	Strongly Disagree	**Disagree**	**Agree**	**Strongly Agree**
Heuristic	0	0	60	40
Normative	0	25	50	25

Tables 7-24 through 7-26 show the results of self-report questions that elicited whether the participants tended to be influenced by feedback and discussion and revised their original recommendations or if they tended to retain their original recommendations. The results are summarized in percentages. The results in Tables 7-24 through 7-26 showed that the both groups tended to be responsive to feedback and discussions.

Table 7-24 shows summary in percentage points of Heuristic and the Normative training group responses to the Likert question that asked if they did not revise their recommendation due to confidence. As shown in Table 7-24, (22%) of the Heuristic training group versus (8%) of the Normative training group strongly disagreed that they did not adjust their recommendation due to confidence. Independent samples *t*-test was used to examine the mean difference between the Heuristic and the Normative training groups on the extent to which they did not adjust their recommendation due to confidence. The Normative training group expressed a higher mean tendency of not revising their recommendation due to confidence (*M = 2.42, SD=.79*) than the Heuristic training group (*M = 2.00, SD = .87*). The Levene's test of equal variance showed insignificant difference of the variation of the Heuristic and the Normative training participants response to the question of if they did not adjust their rating because of confidence *F (1, 20) = .70, p=.41*.The mean difference between the two groups was statistically insignificant at α *=.1*, $|t|$ *(19) =1.15, p = .27.*

Table 7-24: Did Not Adjust Rating Because of Confidence?

	Strongly Disagree	**Disagree**	**Agree**	**Strongly Agree**
Heuristic	22.2	66.7	0	11.1
Normative	8.3	50	33.33	8.3

Table 7-25 shows summary in percentage points of the Heuristic and the Normative training group responses to the Likert question that asked if they did not revise their recommendation because they did not want to use others ideas. Independent samples *t*-test was used to examine the mean difference between the Heuristic and the Normative training groups on the extent to which they did not adjust their recommendation because they didn't want to use others ideas. The Heuristic training group expressed a higher mean of not revising their recommendation because they did not want to use others ideas (*M = 1.67, SD = .71*) than the Normative training group (*M = 1.50, SD = .67*). The Levene's test of equal variance showed insignificant difference of the variation of the Heuristic and the Normative training participants response to the question of if they did not adjust their recommendation because they did not want to use others idea *F (1, 20) = .01, p=.95*. The mean difference between the two groups was statistically insignificant at $\alpha = .1$, $|t|\,(19) = .55, p = .59$.

Table 7-25: Did Not Adjust Rating Because Did Not Want to Use Others Ideas?

	Strongly Disagree	**Disagree**	**Agree**	**Strongly Agree**
Heuristic	44.4	44.4	11.1	0
Normative	58.33	33.33	8.3	0

Table 7-26 shows summary in percentage points of Heuristic and the Normative training participant's responses to the Likert question that asked if they did not revise their recommendation because they did not learn from discussion. Independent samples *t*-test was used to examine the mean difference between the Heuristic and the Normative training groups on the extent to which they did not adjust their recommendation because they did not learn from discussion. The Heuristic training group expressed a higher mean tendency of not revising their recommendation because they did not learn from the discussion (*M = 1.56, SD = .73*) than the

Normative training group (*M = 1.33, SD =.49*). The Levene's test of equal variance showed insignificant difference of the variation of the Heuristic and the Normative training participants response to the question of whether they did not adjust their recommendation because they did not learn from discussion *F (1, 20) = 2.68, p=.12*.The mean difference between the two groups was statistically insignificant at α *=.1,* $|t|$ *(19) =.84, p = .41*.

Table 7-26: Did Not Adjust Rating Because Did Not Learn From Discussions?

	Strongly Disagree	**Disagree**	**Agree**	**Strongly Agree**
Heuristic	55.6	33.3	11.1	0
Normative	66.7	33.33	0	0

7.3.2. Performance on Pre-Requisite Tasks

This section presents the results of analysis that investigated the participants performance on the pre-requisite tasks required for making item difficulty judgments on the probability scale as inferred from the measurement theories and cognitive psychology judgment heuristics research paradigm. The results showed that most of Heuristic training participants scored above average on the practice categorization tasks, of designating items to DOK, content strands, and GLCE. The expectation was that the categorization tasks as opposed to taking the test, would elicit deeper processing of the Practice test items and enhance recall and recognition of test items. However, there was no observed difference between the Heuristic training group and the Normative training group on the measures of their capacity to recall and to recognize similarities between the items on the Real test that appeared on the Practice test.

From Tables 7-27 and 7-28, we can see that both groups were matched in terms of performance on the memory tasks. Both groups had (99%) recall rate of the Practice test items that were replicated on the Real test. The Normative and Heuristic training groups scored .01

apart on the average on the familiarity test, with the Heuristic average being (83%) and the Normative group being (84%). Table 7-27 summarizes the Heuristic participants' performance on the pre-requisite tasks of categorization, recall, and recognition tasks. The results in Table 7-27 are presented in terms of raw scores as well as in proportions correct, with the latter in parenthesis. The results includes the Heuristic training participants' scores on the task of categorizing practice test items in terms of DOK levels, content strands, and third grade level expectations. Each participant's response was scored using content expert designations of the items in the 2005 alignment study conducted by the Michigan Department of Education (MDE). Each participant's responses to the three categorization tasks namely designation of content strand, GLCE, and DOK of items was scored out of the number of items on the test (i.e., 15). A composite, average score on the categorization tasks was computed for each participant and is also presented in Table 7-27.

As can be seen in Table 7-27 the Heuristic training participants' composite scores on the categorization tasks ranged from (40%-87%) percent and yielded a group average score of (69%). Table 7-27 also shows their scores on the questions testing their memory. The questions were asked to: (1) test their capacity to recall previously seen test questions and, (2) recognize items based on similarity of knowledge and skill measured as the Practice test item(s). Except for two participants who could not recall one of the previously presented Practice test question, all could recall the replicated test questions, resulting in the group average recall rate of (99%). The responses to the test of familiarity of the Real test questions were more varied, with an average of (83%) recognition rate for the group as can be seen from Table 7-27.

Table 7-28 summarizes the Normative training participants' performance on the Practice tasks. Presented in Table 7-28 are the participants test score on the fourth grade Practice test.

Their responses to the test items were scored out of 15. Table 7-28 also shows the results of the Normative training participants on the recall and recognition memory test questions.

Table 7-27: Heuristic Group Pre-Requisite Tasks Performance

Participant	Content Strands	GLCE	DOK	Category Score	Recall	Familiarity
1	14(.93)	11(.73)	10(.67)	35(.78)	14 (.93)	12(.80)
2	15(1)	13(.87)	10(.67)	38(0.84)	14 (.93)	13(.87)
3	14(.93)	13(.87)	12(.80)	39(.87)	15 (1)	11(.73)
4	15(1)	0(0)	12(.80)	27(.60)	15 (1)	13(.87)
5	12(.80)	0(0)	7(.47)	19(.42)	15 (1)	14(.93)
6	10(.67)	0(0)	8(.53)	18(.40)	15 (1)	13(.87)
7	14(.93)	13(.87)	8(.53)	35(.78)	15 (1)	11(.73)
8	15(1)	15(1)	7(.47)	37(.82)	15 (1)	12(.80)
9	12(.80)	9(.60)	9(.60)	30(.67)	15 (1)	14(.93)
10	12(.80)	13(.87)	8(.53)	33(.73)	15 (1)	12(.80)

Note: Those that scored 0 on GLCE designation were non respondents; Category score is the pooled average score on Content Strands, GLCE, and DOK and is out of 45; All other scores are out of 15 i.e. the number items on the test.

As is shown in Table 7-28, seven out of twelve of the Normative training participants got perfect score on the Practice test, three answered one question incorrectly, and one participant answered two questions incorrectly. Except for one participant who did not recall one of the previously presented Practice test questions, all the Normative training participants recalled the replicated test questions, giving an overall group average of (99%) recall rate. Just like those of the Heuristic training participants, the responses of the Normative training participants on the test of familiarity of the Real test questions were more varied, with group average of (84%) as can be seen from Table 7-28.

Table 7-28: Normative Group Pre-Requisite Tasks Performance

Participant	Recall	Familiarity	Test Score
1	15 (1)	9 (.60)	15(1)
2	15 (1)	13(.87)	14(1)
3	14 (.93)	6 (.40)	13(.87)
4	15 (1)	14(.93)	15(1)
5	15 (1)	12(.80)	15(1)
6	15 (1)	15(1)	14(.93)
7	15 (1)	13(.87)	15(1)
8	15 (1)	15(1)	15(1)
9	15 (1)	13(.87)	15(1)
10	15 (1)	12(.80)	14(.93)
11	15 (1)	15(1)	15(1)
12	15 (1)	15(1)	14(.93)

Note: # of items on the test = 15; all scores are out of 15

Because the feedback for both training groups served to present them with the empirical item difficulties and the judgments of other participants, rank ordering of the items based on these norms was correlated with each participants rank ordering of the items and statistics obtained for the groups. These statistics were obtained to investigate how the item difficulty rank ordering of the participants correspond with the item difficulty ranks based on the norms across the rounds of judgment. These statistics are also to indicate if monotonicity of the participant's judgment with these norms and variance reduction of their judgments are in play in the feedback rounds of judgment. Tables 7-29 and 7-30 that follow presents the means and the standard deviations of the correlations for the empirical item difficulty ranks and the training study group mean ranks, respectively.

Table 7-29 shows the means and standard deviations of the Spearman rank correlations of participants' rankings of items in terms of difficulty with empirical difficulty rankings of the items based on the empirical proportion of entire fourth graders responding correctly to the items in 2005. In Table 7-29, the column labeled "*M*" contains the means of correlations of the participants' ranking of items in terms of difficulty with the empirical difficulty ranks of the items in 2005 while the column labeled “*SD*” contains the standard deviation of the correlations. The mean correlations for both groups consistently increased while the standard deviations decreased across rounds of judgments. For the Normative training group there was marked increase in the average correlation of the participants item difficulty rankings with the empirical difficulty ranks of the items following construct map feedback (.25 between the first and second round feedback rounds), compared to the Heuristic training group which increased by .07.

Independent samples *t*-tests were obtained to examine the mean difference between the Heuristic and the Normative training groups on the extent to which their item difficulty rankings across rounds of judgment related with the empirical difficulty ranks of the items. In the practice round of judgment on average, the Heuristic training group rank ordering of items correlated higher with the empirical difficulty ranks of the items than the Normative training group. The Levene’s test of equal variance showed insignificant difference of the variation of the correlation between the Heuristic and the Normative training participants rank ordered item difficulties with empirical difficulty rankings of the items in the practice round *F (1, 20) = .08, p=.78.* The mean difference between the two groups was statistically insignificant at α *=.1,* $|t|$ *(20) =.18, p* =.86.

In the first feedback round of judgment, the Normative training group rank ordering of items on the average correlated higher with the empirical difficulty ranks of the items than the Heuristic training group. The Levene’s test of equal variance showed insignificant difference of the variation of the correlation between the Heuristic and the Normative participants rank ordering of items with the empirical item difficulty ranks *F (1, 20) = .80, p=.38.*The mean difference between the two groups was statistically insignificant at α *=.1,* $|t|$ *(20) =.77, p* =.45.

In the second feedback round of judgment that followed construct map feedback for the Normative training group, on average the Normative training group’s rank ordering of items correlated higher with the empirical difficulty ranks of the items. The Levene's test of equal variance showed insignificant difference of the variation of the Heuristic and the Normative training groups rank ordering of items correlations with the empirical difficulty ranks of the items *F (1, 17) =2.38, p=.14*. The mean difference between the two groups was statistically significant at α *=.1,* $|t|$ *(17) =6.31, p* <.001.

Table 7-29: Means and Standard Deviations of the Correlations of Judged With Empirical Difficulty Ranking of the Items

Group	**Round**	***M***	***SD***
Heuristic	Practice	.52	.27
Heuristic	Round One	.62	.12
Heuristic	Round Two	.69	.10
Normative	Practice	.50	.24
Normative	Round One	.67	.17
Normative	Round Two	.92	.04

Table 7-30 shows the means and standard deviations of the Spearman rank correlations of participants' rankings of items in terms of difficulty with their study group mean item difficulty rankings. In the practice round of judgment on average, the Normative training group rank ordering of items correlated higher with their study group mean item difficulty ranks of the items than the Heuristic training group. The Levene's test of equal variance showed insignificant difference of the variation of the correlation between the Heuristic and the Normative training participants rank ordered item difficulties with the study group mean item difficulty rankings of the items in the practice round *F (1, 20) = .19, p=.67.* The mean difference between the two groups was statistically insignificant at α *=.1,* $|t|$ *(20) =1.10, p* =.29.

In the first feedback round of judgment on average, the Normative training group rank ordering of items correlated higher with the study group mean difficulty ranks of the items than the Heuristic training group's. The Levene's test of equal variance showed insignificant difference of the variation of the correlation between the Heuristic and the Normative participants rank ordering of items with the study group mean item difficulty ranks *F (1, 20) = .10, p=.75*.The mean difference between the two groups was statistically insignificant at α *=.1,* $|t|$ *(20) =.93, p* =.37. In the second feedback round of judgment that followed the construct map feedback for the Normative training group, on average the Normative training group's rank ordering of items correlated higher with their study group mean item difficulty ranks than the Heuristic training group. The Levene's test of equal variance showed significant difference of the variation of the Heuristic and the Normative training participants item difficulty rank ordering with their study group mean item difficulty ordering *F (1, 17) =5.69, p=.03*. The mean difference between the two groups was statistically significant at α *=.1,* $|t|$ *(17) =6.55, p* <.001.

Table 7-30: Means and Standard Deviations of the Correlations of Individual Judged With Study Group Mean Difficulty Ranking of Items

Group	**Round**	***M***	***SD***
Heuristic	Practice	.62	.20
Heuristic	Round One	.68	.12
Heuristic	Round Two	.74	.09
Normative	Practice	.71	.19
Normative	Round One	.74	.18
Normative	Round Two	.93	.03

7.3.3. Exploratory Check of Judgments for Fit With the Heuristic Model

This section presents the results of exploration of the primary data of this dissertation, namely the item difficulty judgments of the participants. The rows of the primary data matrix comprised of each participants and the columns of each items difficulty judgment, respectively. The results presented in this section are evidence pertaining to qualitative check of assumptions made about the knowledge and skills constructs of the PLD stimulus, measured by the test items, and about the cognitive processes underlying participant's judgment.

The evidence presented in this sub-section is also consistent with Kane's (2001) internal validity criterion and Raymond and Reid's (2001) criterion that the judgments of a well-trained participant should conform to assumptions of the standard setting method including, about the standard setting materials, participants, and activities. The analytic technique used to check tenability of the assumptions is the principal coordinates (aka metric or classical scaling) procedure. The first evidence presented is the principal coordinates dimensionality reduction analysis checks of whether the judgment datasets can be summarized in relatively few items and persons knowledge and skills constructs dimensions. Since the results of this first evidential check, suggested that it makes sense to summarize the data in a few dimensions, the second evidence presented is scatter plots obtained to check for patterns of clustering of the participants and items in a two-dimensional space. The specific results of the principal coordinates analysis dimension reduction is presented in the discussion that immediately follow. The results showed that the eigenvalues associated with the first principal coordinate explained a substantial chunk of the variation of the rows and columns of the matrix of judged item difficulties over rounds of recommendations.

Figure 7-1 and Figure 7-2 shows scree plot graphs of eigenvalues of the row dimensionality reduction analysis plotted against the principal coordinate number, for the Heuristic and the Normative training group data, respectively and for the rounds of recommendations. In Figure's 7-1 and 7-2, the blue line is the plot of eigenvalues for the practice round, the red for feedback round one, and the green for feedback round two of judgment data matrix. The elbow of the scree plots all appeared on the second principal coordinate. In Figure 7-1 for the Heuristic training group data, the eigenvalue of the first principal coordinates of the row dimension data explained (44%) in the practice round, (35%) in feedback round one, and

(43%) in feedback round two of the variance in the data. The eigenvalues of the first two principal coordinates explained (64%) in the practice round, (67%) in feedback round one, and (66%) in feedback round two of the variance in the data. In Figure 7-2 for the Normative training group data, the eigenvalue of the first principal coordinates of the row dimension data explained (41%) in the practice round, (31%) in feedback round one, and (41%) in feedback round two of the variance in the data. The eigenvalues of the first two principal coordinates explained (58%) in the practice round, (57%) in feedback round one, and (58%) in feedback round two of the variance in the data. The scree plot shows shrinking of the variation in the judged item difficulties for both groups row dimension data across rounds of recommendation. These findings indicated that it makes sense to summarize the row dimension data set for the Heuristic and the Normative training groups in relatively few dimensions.

Figure 7-1: Scree-Plots of Eigenvalues of the Heuristic Group Row Dimension Reduction (For interpretation of the references to color in this and all other figures, the reader is referred to the electronic version of this dissertation)

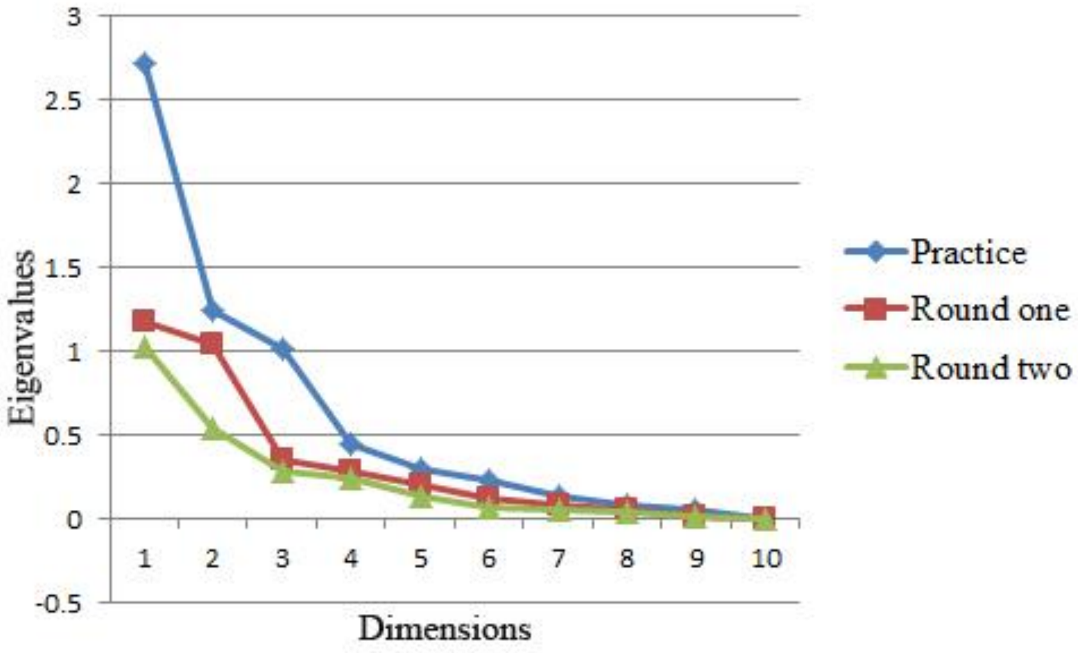


Figure 7-2: Scree-Plots of Eigenvalues of the Normative Group Row Dimension Reduction

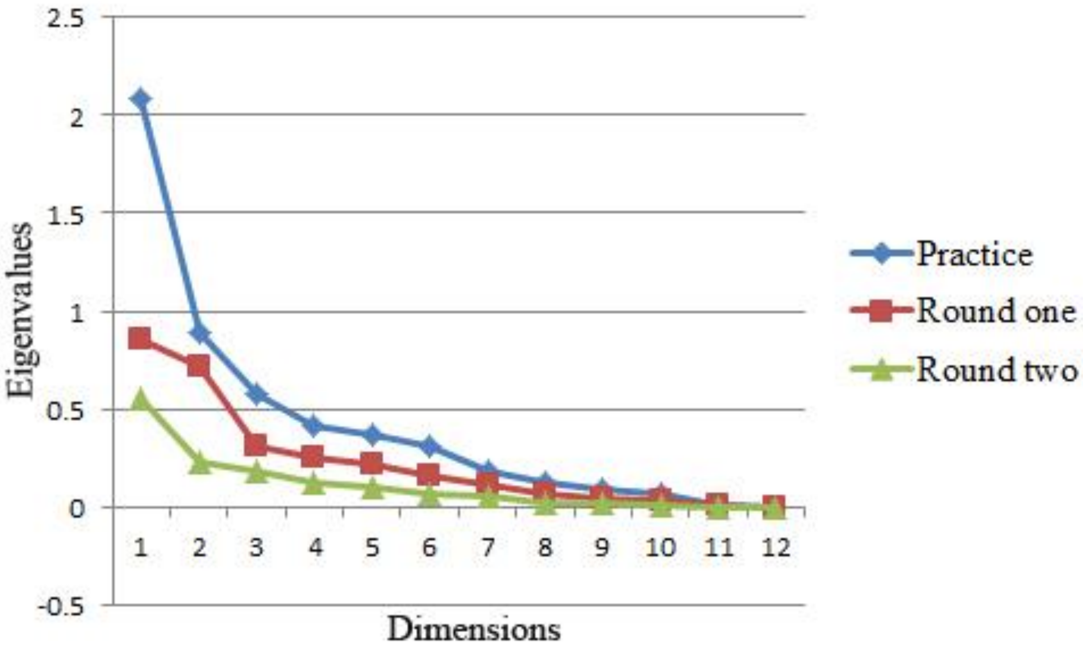


Figure 7-3 and Figure 7-4 are scree-plots of eigenvalues for Heuristic and Normative training group columns principal coordinates dimension reduction, respectively. Figure 7-3 and Figure 7-4 shows that the elbow of the scree plots across rounds, for both groups, occurred at the second principal coordinate. In Figure's 7-3 and 7-4, the blue line is the plot of eigenvalue for the practice round, the red for feedback round one, and the green for feedback round two of judgment data matrix. The elbow of the scree plots all appeared on the second principal coordinate.

In Figure 7-3 for the Heuristic training group, the eigenvalue of the first principal coordinate of the column dimension explained (53%) in the practice round, (60%) in feedback round one, and (68%) in feedback round two of the variance in the data. The eigenvalues of the first two principal coordinates explained (71%) in the practice round, (77%) in feedback round one, and (82%) in feedback round two of the variance in the data.

In Figure 7-4 for the Normative training group, the eigenvalue of the first principal coordinate of the column dimension explained (70%) in the practice round, (64%) in feedback round one, and (86%) in feedback round two of the variance in the data. The eigenvalues of the first two principal coordinates explained (80%) in the practice round, (77%) in feedback round one, and (90%) in feedback round two of the variance in the data.

These findings indicated that it makes sense to summarize the column dimensions for both the Heuristic and the Normative training group data set in relatively few dimensions.

Figure 7-3: Scree-Plots of Eigenvalues of the Heuristic Group Column Dimension Reduction

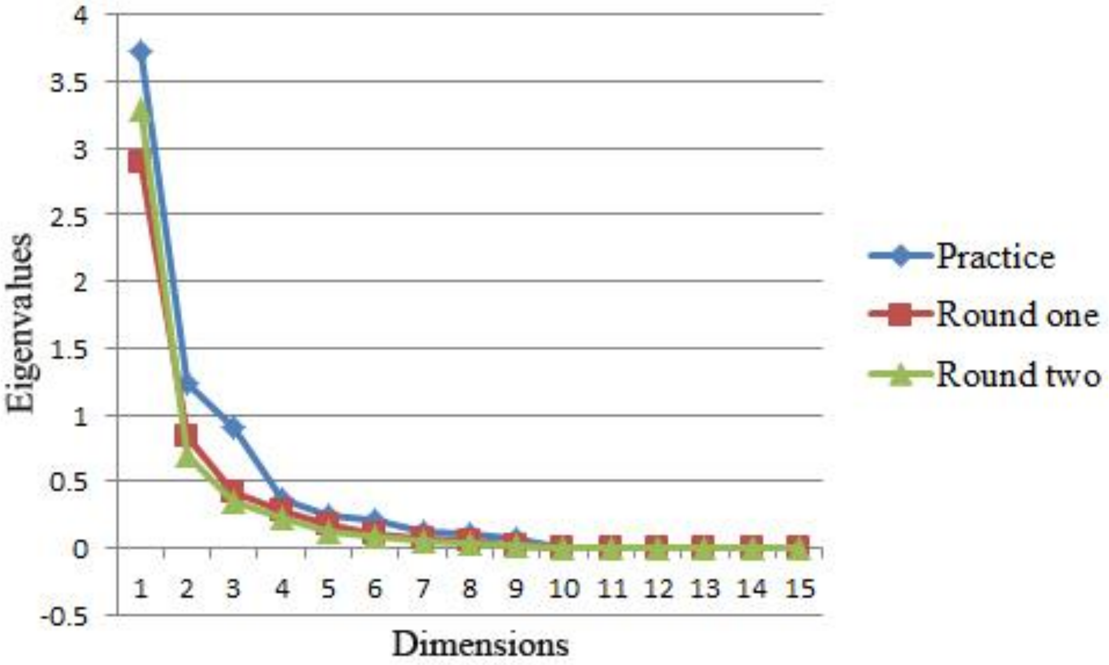
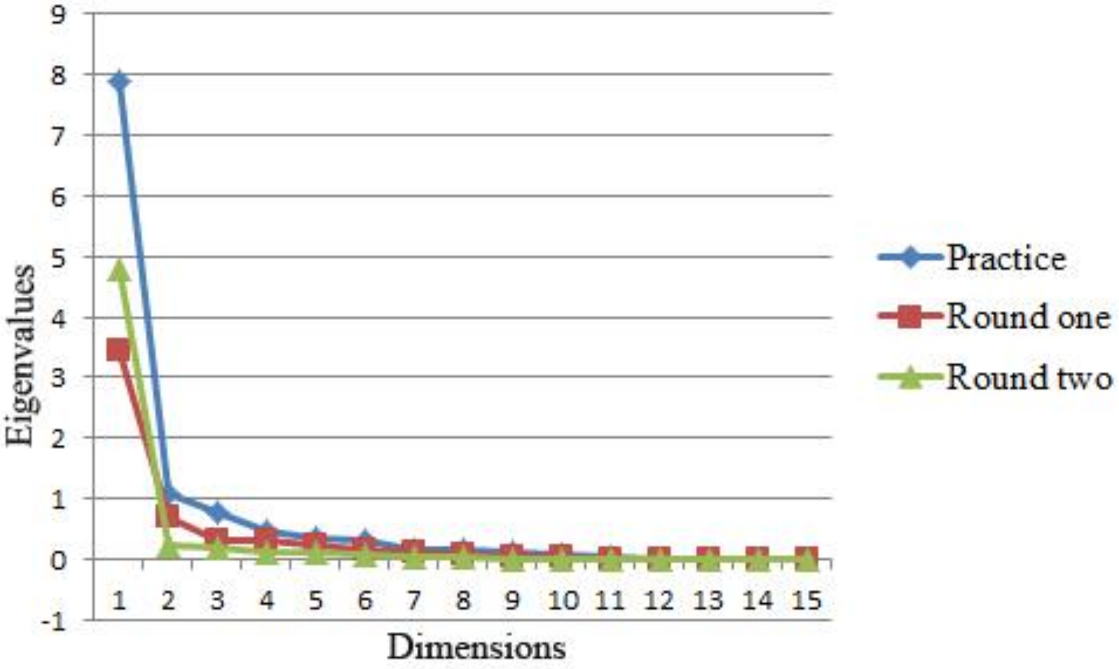


Figure 7-4: Scree-Plots of Eigenvalues of the Normative Group Column Dimension Reduction



In addition to Figures 7-1 through 7-4 that explored the modality and dimensionality of the knowledge and skills constructs of barely proficient performance and measured by the tests, respectively, based on the judged item difficulty data, scatter plots were obtained also, to visualize how the participants and items positioned on the first two principal coordinates and to explore the factors that explain similarities. These plots also were meant to investigate whether the heuristic model assumptions that recall of experienced members of the student performance category, that categorization of items in terms of knowledge and skills they measure, and recall of information about experienced items categories mediates items difficulty judgments were tenable.

Because of interest in this dissertation is about the factors that drive participants' item difficulty judgments, the principal coordinates scatter plot data summary presented here are only for the column dimension of the data matrix. However, the interested reader should please refer to appendix D for the scatter plots for the principal coordinates for the row dimension reduction showing positioning of participants across rounds. Meanwhile, the latter scatterplots showed that in the practice round that the participants tended to cluster together according to indicators of teaching experience and math specialization. However the observed clustering by experiential factors became less apparent in rounds one and two, instead proximity tended to depend on table group allocation and also their appeared to be overall group relatedness.

Figures 7-5 and 7-6 that immediately follow shows scatter plots of the first two principal coordinates of the column dimension reduction for the Heuristic and the Normative training groups, respectively for their practice round recommendation with item DOK and Content Strand point labels. The first code of the point labels stands for the content strand and the second for the

DOK level of the item (Content Strands: N = Number and Operations; D = Data and Operations; G = Geometry; M = Measurement).

The scatter plots of the first two principal coordinates of the column dimension reduction of the participants' judged item difficulties shows items of the same content strand and DOK being more proximal in the two-dimensional space. It also shows similarity of some sort of the scale values of the items of the same DOK and difference between items of different DOK levels along the first principal coordinate. The GLCE's of the items were also provided to the participants as part of the training. However, the finding from the plots in which the GLCE's and content domains were inserted was that items did not cluster in terms of GLCE and Domains measured by the items.

Figure 7-5: Plot of the Heuristic Group Practice Round Principal Coordinates With Item Content Strand and DOK Point Labels

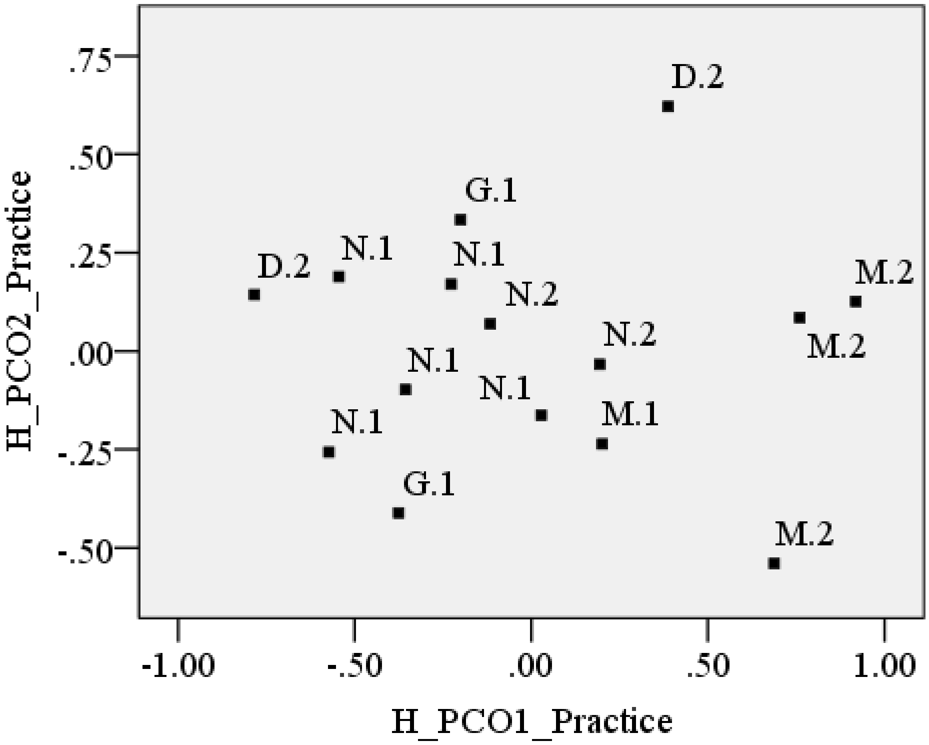
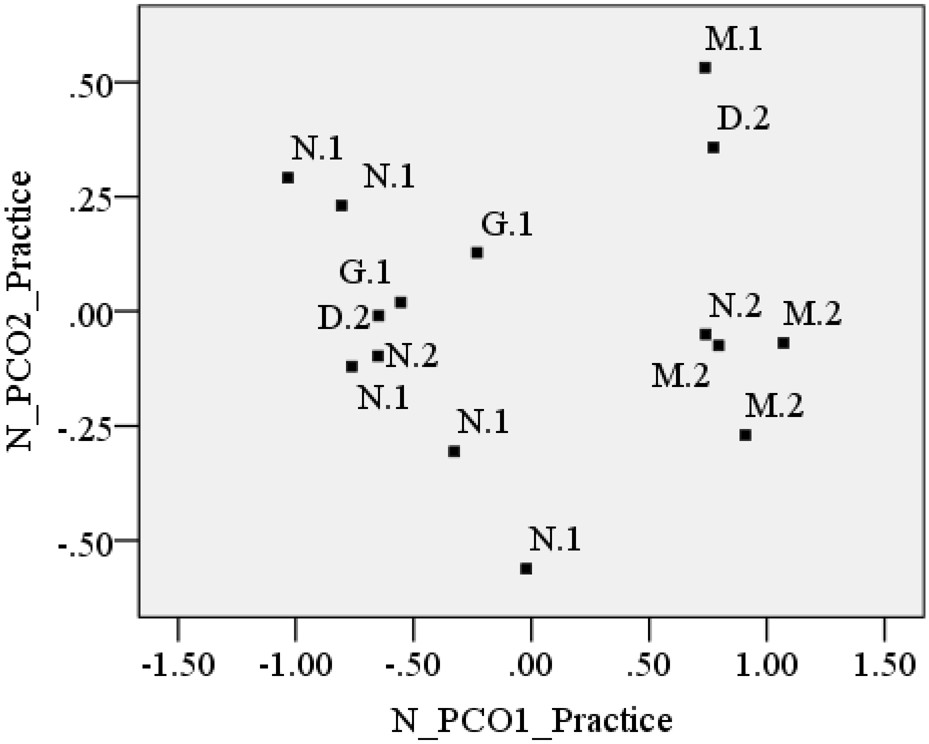


Figure 7-6: Plot of the Normative Group Practice Round Principal Coordinates With Item Content Strand and DOK Point Labels



Figures 7-7 and 7-8 that immediately follow shows the scatter plots of the first two principal coordinates of the column dimension reduction for the Heuristic and the Normative training groups, respectively, feedback rounds one and two recommendations with item DOK and Content Strand point labels inserted. The first code of the point labels stands for the content strand and the second for the DOK level of the item (Content Strands: N = Number and Operations; D = Data and Operations; G = Geometry; M = Measurement).

In the scatterplot in Figure 7-7 for the Heuristic training group feedback rounds one and two judgments, just like those of practice, showed items clustering by depth of knowledge levels along the first principal coordinate. However, compared to the practice round, the clustering pattern in terms of content strands and DOK became less prominent. In the scatterplot in Figure 7-8 for the Normative training group feedback rounds one and two judgments, the clustering of items by content strands and depth of knowledge level compared to those of the Heuristic training group became considerably less prominent.

Figure 7-7: Plot of the Heuristic Group Feedback Rounds One and Two Principal Coordinates With Content Strand and DOK Point Labels

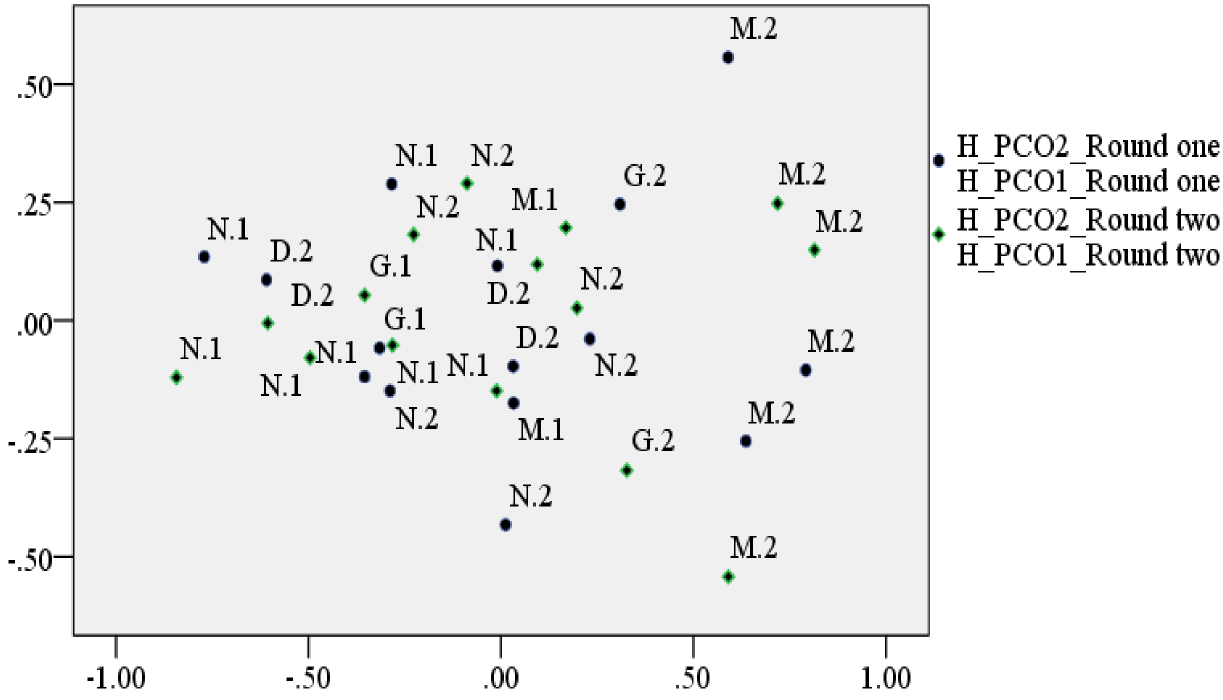
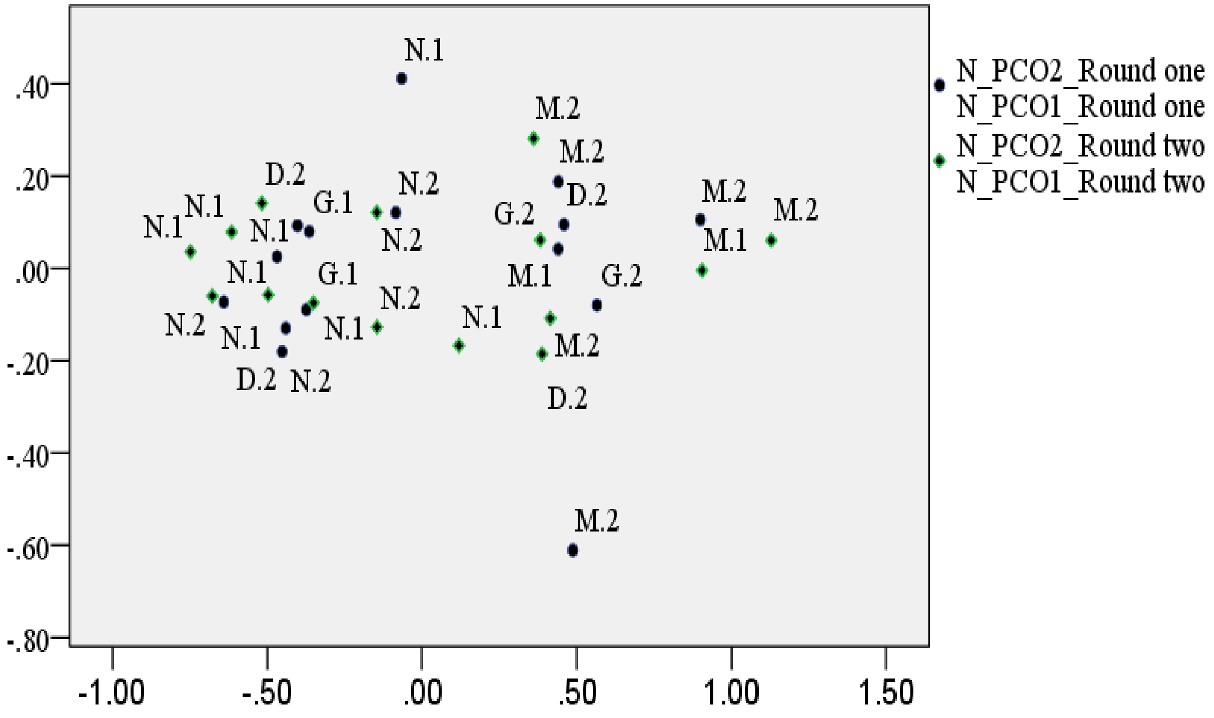


Figure 7-8: Plot of the Normative Group Feedback Rounds One and Two Principal Coordinates With Content Strand and DOK Point Labels



7.3.4. Correlations of Judgments With Other Measures

The evidence presented in the preceding section suggested that it makes sense to summarize the judgment data set of the Heuristic and the Normative training groups in terms of a single dimension of item difficulty, therefore in this section the evidence presented are about the test of the assumption that the Heuristic training would result in item difficulty judgments that have higher technical qualities of reliability and validity.

The comparative reliability of the judgments of the Heuristic and the Normative training groups' was investigated by the Mantel random permutation test of significance of the correlations of the Euclidean distance matrices of the rounds of judgment data of the items replicated on both the Practice and the Real tests. This Mantel random permutation test of the significance of difference from zero of the Euclidean distance matrices of the replicated items correlations is consistent with the Raymond and Reid's framework (2001) stability of judgments across occasions criterion measure for evaluating the judgments of well-trained standard setting participants.

The comparative construct validity of the judgments of the Heuristic and the Normative training groups was evaluated by correlating participant's judgments with other criterion measures of item difficulties and via independent sample *t*-tests of significance of the difference between the average correlations of the two training methods. The four criterion measures that were used to evaluate participants judgments for construct validity were as follows: the modal DOK designation of the items by content experts that participated in MEAP alignment study in 2005, empirical proportion of entire fourth grade students that responded correctly to the items in 2005, participants training group mean item difficulty estimates, and the item difficulties estimated at the bootstrap estimated PLD cut score[14]. Therefore, the correlational measures obtained to investigate the technical qualities of reliability and validity of participant's judgments in the order presented were as follows:

(1) Correlation between Euclidean distance matrices of the rounds of judgment data of the items that were replicated on both the Practice and the Real tests;

(2) Correlation of participants estimated item difficulties for the barely proficient with the modal DOK designation of the items by content experts that participated in MEAP alignment study in 2005;

(3) Correlation of participants estimated item difficulties for the barely proficient with the empirical item difficulties for entire fourth grade students based on student responses to the items in 2005 MEAP test;

(4) Correlation of participants estimated item difficulties for the barely proficient with their study group mean item difficulty estimates;

(5) Correlation of participants estimated item difficulties for the barely proficient with the item difficulties estimated at the bootstrap PLD cut score estimate.

Before presenting the correlational evidences, it is deemed appropriate to discuss first the bootstrap PLD cut score and item difficulty estimation approach which were used to cross validate participant's judgments. The PLD bootstrap cut score estimation approach was based on criterion referencing to the PLD knowledge and skills descriptions. The resulting cut score estimated based on this PLD bootstrap cut score estimation approach was to facilitate evidence that is consistent with the Raymond and Reid's (2001) criterion of reflection of realistic

[14] Please note that only items of DOK levels 1 and 2 were on the tests so that the modal content expert DOK designation of the items variable is a binary variable

expectations. The MEAP 2005 fourth grade math Proficient performance PLD was operationalized using the bootstrap resampling technique to generate the best guess estimate of the ability of barely proficient using the principles of Rasch item response model and the central limit theorem.

Hence, the bootstrap PLD cut score although computed mechanistically, however was based on criterion-referencing to the knowledge and skills description of the PLD. The assumption made to facilitate operationalizing the PLD for the task of evaluating the quality of the outcomes, was that the PLD is an adequate model of the knowledge and skills of the entire proficient student population. Consequently, threshold mastery of the knowledge and skills of the MEAP 2005 fourth grade math Proficient PLD was operationalized as a 50:50 chance of responding correctly to exemplar items measuring them. The hypothesized Proficient cut score is the average of the sampling distribution of the mean of Rasch model estimated difficulties of the subset fourth grade released MEAP math items that the researcher judged to measure the knowledge and skills specified in the Proficient PLD. Item difficulties on the probability scale were estimated for each item on the Practice and the Real test by plugging into the Rasch model, the bootstrap mean estimated cut score, for the latent ability, and the Rasch model calibrated difficulty estimate of the items in 2005 for item difficulty. Consequently, the hypothesized cut score on the test score scale (aka true score scale) for the Practice and the Real tests was the sum of these difficulty estimates for the items. To summarize, the general steps used in generating the hypothetical PLD cut score and mapping it to the Practice and the Real test cut scores were as follows:

(1) Match each knowledge and skills descriptor of the PLD to available items that measure them

(2) Operationalize the items in abstract, in terms of their IRT model calibrated difficulties

(3) Compile the difficulties of all available items that measure the knowledge and skills descriptors of the PLD

(4) Consider the difficulties of the sampled items measuring the knowledge and skills of the PLD, as representative of the population distribution of the abilities that are barely located at the student performance category of the PLD

(5) Take 1000 bootstrap samples, each of size of the sampled item difficulties measuring the knowledge and skills of the PLD

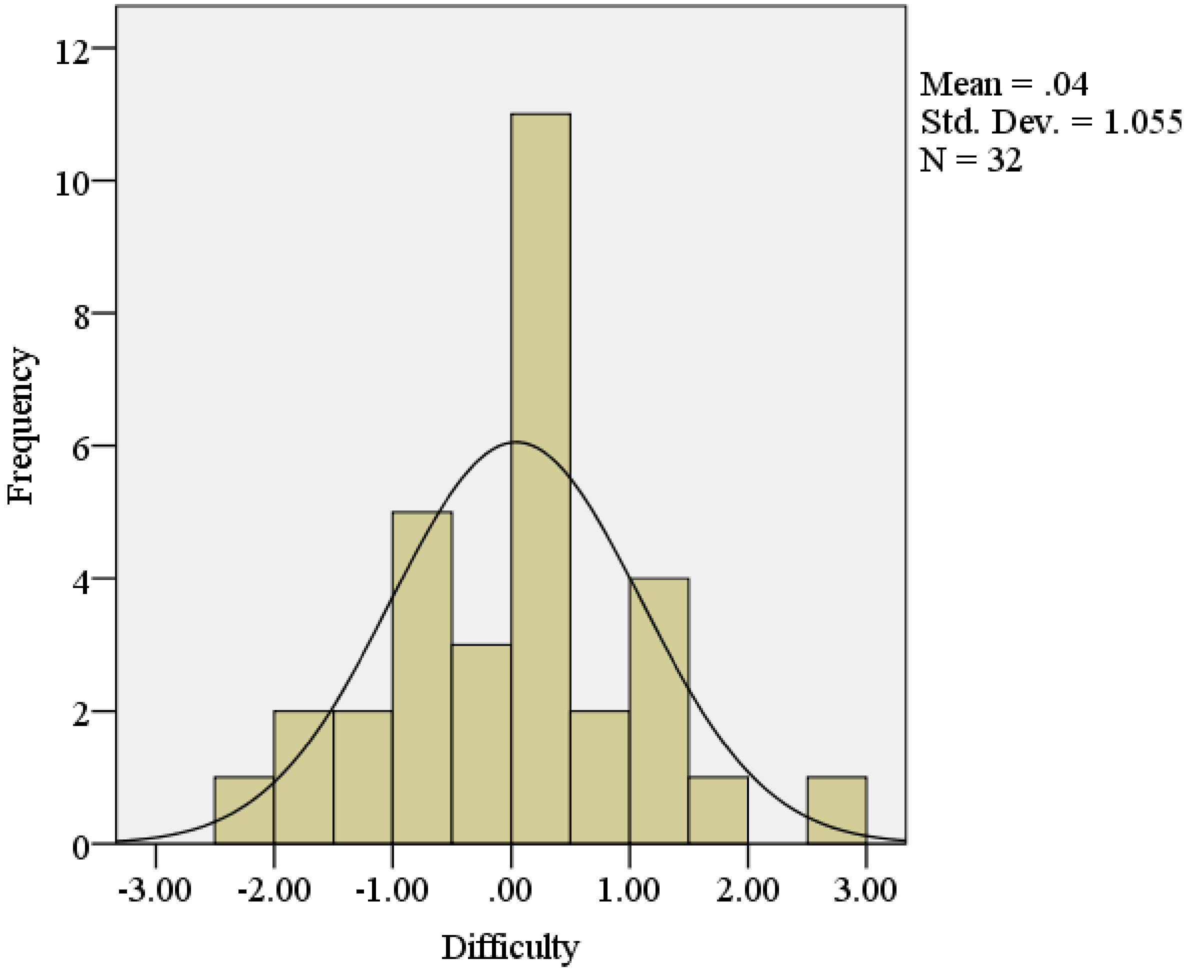
(6) Compute the mean and the standard deviation of each of the 1000 bootstrap samples of item difficulties

(7) The mean of the 1000 mean difficulties is the cut score on the IRT ability scale while the standard deviation is its standard error

(8) The cut score or the true score on the test scale, for the modified Angoff procedure is the sum of the probabilities of correct response of the items on the test computed at this IRT ability and using their IRT model calibrated difficulties.

Figure 7-9 shows histogram plotting the Rasch model difficulties of the released 2005 MEAP items ($N = 32$), judged by the researcher as measuring the knowledge and skills delineated in the math Proficient PLD with normal curve overlaid. Figure 7-9 shows clearly a unimodal distribution of item difficulties of this subset items.

Figure 7-9: Histogram of the Rasch Model Difficulties of Items Measuring the MEAP 2005 Fourth Grade Math Proficient PLD



The cut score estimate was obtained as bootstrap resampling mean of Rasch model difficulties of the selected subset items measuring the knowledge and skills of the PLD. Table 7-31 shows the summary statistics obtained based on the bootstrap resampling procedure. The bootstrap resampling was based on drawing 1000 samples. The confidence interval type was the percentile and a 95% confidence interval. The PLD cut score on the IRT ability scale was, *Cut score = .04* with confidence bound [-.33, .38] (see the Table 7-31 for other details).

Table 7-31: Bootstrap Resampling PLD Cut Score Statistics

Statistic	Sample Size	Std. Error	95% Confidence Interval Lower	95% Confidence Interval Upper
.04	32	.18	-.33	.38

Note: bootstrap results are based on 1000 bootstrap samples

The Practice and the Real test items difficulties on the proportion correct scale for the barely proficient were estimated at the bootstrap PLD *(cut score = .04)* by plugging into the Rasch IRT model the cut score as the latent ability and the item difficulties on the IRT ability scale, and these were correlated with the participants judgments which results for the training groups would be summarized subsequently.

The item difficulties for the barely proficient estimated for the subsets of items on the Practice and the Real test were summed to obtain the cut score on the true score scale for the Practice and the Real tests, respectively as presented in Table 7-32 that follows. Also, the difficulties of items on the Practice and the Real test were computed at the IRT ability end points of the bootstrap PLD cut score confidence interval. The estimated conditional difficulties of the items at the left and right end of the ability confidence interval were each summed for the subset of items on the Practice and the Real test to obtain the confidence bounds on the true score scale for the Practice and the Real test cut scores. The estimated cut score on the true score scale computed based on the above process for the Practice test was 8.43 and 8.15 for the Real test. Please refer to Table 7-32 for other details.

A little bit of discussion to rationalize the use of the PLD bootstrap derived cut scores and the derived item difficulty estimates for the barely proficient for evaluating the participants judgment outcomes are as follows: First, it is worthy of note that to the extent that the selected items measuring the PLD are representative, the bootstrap cut score and the item difficulty estimates for the barely proficient are precise and approximate the true values. Second, substantively the bootstrap estimates were reasonable because the cut score for the Real test is lower than that for the Practice test and it was supposed to be more difficult at least in terms of the substantive domain constructs being measured by the tests. In accordance with the content expert designations of DOK, there were more items of DOK level 2 on the Real test. In terms of specifics, there were seven items on the Practice test designated by content experts at DOK level of 2 and nine items on the Real test designated at DOK level of 2. Hence, it makes both substantive and technical sense that the cut score should be lower for a more difficult test than an easier test. Therefore, it makes sense that the cut score is lower for the Real test than the Practice test. Also, in measurement technical terms the Practice and the Real tests are essentially parallel forms because as shown in Table 7-32 the confidence bounds around their cut score overlaps. As shown in the Table 7-32 the cut score for the Practice test is 8.43 and that for the Real test 8.15. Consequently, based on the foregoing observations it was considered appropriate to use the bootstrap estimates as the standard for cross validating the participant estimates for reasonableness.

Table 7-32: Proficient PLD True Score Scale Bootstrap Cut Score Estimates.

Test	**Cut Score Estimate on True Score Scale**	**Lower Confidence Bound on True Score Scale**	**Upper Confidence Bound on True Score Scale**
Practice	8.43	7.27	9.45
Real	8.15	7	9.18

The rest of the discussion in this section presents correlational evidence about the technical quality of the participant’s item difficulty judgments. The correlational evidence presented includes those of Euclidean distances matrices of the rounds of judgments on the items that were replicated on both the Practice and the Real tests and based on relating participant’s judgment to the aforementioned four criterion measures of item difficulty. The evidence is presented first for those of the Euclidean distance matrices which addressed the reliability of the participants judgments, and followed by those based on relating the participant's judgment to other measures of item difficulty with the latter addressing the construct validity of the participant's judgments.

Table 7-33 shows the Pearson correlations between the Euclidean distance matrices of the subset of the judgment data sets of the items that were replicated on the Practice and the Real test. Table 7-33 also presents the random permutation test of significance of the correlations. The permutation test of the significance of the correlation between matrices of Euclidean distances is called the Mantel test. The result shows positive relationships of the distances between the replicated items over rounds of recommendation.

For the Heuristic training group all the correlations between the rounds of judgment distance matrices of the replicated items remained, high, fairly stable, and significantly different from 0. The Normative training group correlations in comparison to those of the Heuristic training group were lower and the correlation between the practice and the feedback round two distance matrices of the replicated items was moderate (.50) and statistically insignificantly different from 0 at $\alpha = .05$.

Table 7-33: Correlations of Euclidean Distances Between Replicated Items

Study Group	**Distance Pair**	**Correlation (*r*)**	***p*-value**
Heuristic	Practice and Round One	0.89	0.02
Heuristic	Practice and Round Two	0.93	0.01
Heuristic	Round One and Round Two	0.95	0.001
Normative	Practice and Round One	0.78	0.001
Normative	Practice and Round Two	0.50	0.07
Normative	Round One and Round Two	0.87	0.04

Next in the line of evidence are the means and standard deviation of the correlations of the participant's judgment with other criterion measures of item difficulty. Tables 7-34 through 7-37 present the means and standard deviations of the Pearson correlations of participant's judgment with the other measures of item difficulties by training method. In Tables 7-34 through 7-37, the column labeled "*M"* is the average of the correlations of the participant's judgment with the involved criterion measure while the column labeled "*SD*" is the standard deviation of the correlations of the participant's judgment with the involved criterion measure. The observations made from Tables 7-34 through 7-37 displaying the means and standard deviations of the correlations of participants item difficulty judgments with other criterion measures of difficulty was that on the average that the participants' estimated item difficulty estimates for the barely proficient tended to correlate higher with their study group pooled item difficulty estimates, followed in order by the PLD cut score estimated item difficulties, the empirically based item difficulties for the entire fourth grade students, then the DOK designation of the items.

Table 7-34 shows the means and standard deviations of the Pearson correlations of participants' judged item difficulties for the barely proficient fourth grader with the modal content expert DOK designation of the items in 2005. As can be observed from Table 7-34 the average correlations between participants' judged item difficulties and content expert DOK designation of items were negative as expected. The latter is because of the negative relationship between the operational definition of item difficulty in terms of proportion of correct response and that of the concept of DOK[15]. In the practice round of judgment on the average, the Normative training group item difficulty estimates correlated higher with the modal DOK assigned to the items by the content experts than the Heuristic training group. The Levene's test of equal variance showed insignificant difference of the variation of the correlations of the Heuristic and the Normative training participants judged item difficulties with the DOK designations of the items in the practice round *F (1, 20) = .26, p=.62.* The mean difference between the two groups was also statistically insignificant at α *=.1,* $|t|$ *(20) =.29, p* =.78.

In the first feedback round of judgment on average, the Heuristic training group item difficulty estimates on average correlated higher with the modal DOK assigned to the items by content experts than the Normative training group. The Levene's test of equal variance showed insignificant difference of the variation of the correlations of the Heuristic and the Normative training participants judged item difficulties with the DOK designations of the items in the first feedback round of judgment *F (1, 20) = .58, p=.46.* The mean difference between the two groups was statistically insignificant at α *=.1,* $|t|$ *(20) =1.08, p* =.29. In the second feedback round of

[15] The depth of knowledge (DOK) is also referred to as the substantive domain item difficulty (see Haertel & Lorie, 2004)

judgment, the Levene's test of equal variance showed insignificant difference of the variation of the correlations of the Heuristic and the Normative training participants judged item difficulties with the modal DOK designations of the items, *F (1, 20) = .32, p=.58.*On the average the Heuristic training group's item difficulty estimates correlated significantly higher with the modal DOK assigned to the items by content experts than the Normative training group's, α *=.1,* $|t|$ *(20) =3.28, p =.004.*

Table 7-34: Means and Standard Deviations of the Correlations of Item Difficulty Judgments With Designated Depth of Knowledge Level of Items

Group	**Round**	***M***	***SD***
Heuristic	Practice	-.41	.19
Heuristic	Round One	-.38	.15
Heuristic	Round Two	-.41	.13
Normative	Practice	-.43	.16
Normative	Round One	-.31	.18
Normative	Round Two	-.25	.10

Table 7-35 shows the means and standard deviations of the Pearson correlations of participants' judged item difficulties with the empirical item difficulties based on entire fourth graders responses. The results shows that on average the correlations tended to increase while the standard deviations of the correlations tended to decrease across rounds of judgment for both groups. In the practice round of judgment , the Levene's test of equal variance showed insignificant difference of the variation of the correlations of the Heuristic and the Normative training participants judged item difficulties with the empirical item difficulties *F (1, 20) = .02, p=.91.* On average, the Normative training group item difficulty estimates correlated higher but statistically insignificantly so with the empirical item difficulties than the Heuristic training group, at α *=.1,* $|t|$ *(20) =.46, p =*.65.

In the first feedback round of judgment on the average, the Normative training group item difficulty estimates correlated higher with the empirical item difficulties than the Heuristic training group. The Levene's test of equal variance showed insignificant difference of the variation of the correlations of the Heuristic and the Normative training participants judged item difficulties with the empirical item difficulties *F (1, 20) = 2.51, p=.13.* The mean difference between the two groups was statistically insignificant at α *=.1,* $|t|$ *(20) =1.69, p* =.11 albeit was nearing significance. In the second feedback round of judgment, that followed construct map feedback for the Normative participants, the Levene's test of equal variance showed insignificant difference of the variation of the correlations of the Heuristic and the Normative training participants judged item difficulties with the empirical item difficulties *F (1, 20) = 2.53, p=.13.* On the average the Normative training group item difficulty estimates correlated significantly higher with the empirical item difficulties than the Heuristic training group at α *=.1,* $|t|$ *(20)=7.54, p<.001.*

Table 7-35: Means and Standard Deviations of the Correlations of Judged With the Empirical Difficulty of Items

Group	**Round**	***M***	***SD***
Heuristic	Practice	.49	.24
Heuristic	Round One	.59	.15
Heuristic	Round Two	.67	.08
Normative	Practice	.54	.20
Normative	Round One	.68	.10
Normative	Round Two	.89	.06

Table 7-36 shows the means and standard deviations of the correlations of participants' judged item difficulties for the barely proficient student with their training group mean item difficulties. The results also shows that the average of correlations tended to increase while standard deviations decreased over rounds for both groups. In the practice round of judgment on average, the Normative training group item difficulty estimates correlated higher with their group mean item difficulty estimates than the Heuristic training group. The Levene's test of equal variance showed insignificant difference of the variation of the correlations of the Heuristic and the Normative training participants judged item difficulties with their training group mean $F\ (1, 20) = 1.24, p=.28$. The mean difference between the two groups was statistically insignificant at $\alpha = .1$, $|t|\ (20) = 1.26, p = .22$.

In the first feedback round of judgment on average, the Normative training group's item difficulty estimates correlated higher with their group mean item difficulty estimates than the Heuristic training group. The Levene's test of equal variance showed insignificant difference of the variation of the correlations of the Heuristic and the Normative training participants judged item difficulties with their study group mean $F\ (1, 20) = 1.86, p=.19$. The mean difference between the two groups was statistically insignificant at $\alpha = .1$, $|t|\ (20) = 1.68, p = .11$ although nearing significance. In the second feedback round of judgment, that followed the construct map feedback for the Normative training group, the Levene's test of equal variance showed significant difference of the variation of the correlations of the Heuristic and the Normative training participants judged item difficulties with their study group mean at $\alpha = .1, F\ (1, 14.69) = 3.47, p=.08$. On the average the Normative training group's item difficulty estimates correlated significantly higher with their training group mean item difficulties than the Heuristic training group, $\alpha = .1$, $|t|\ (14.69)=6.85, p<.001$.

Table 7-36: Means and Standard Deviations of the Correlations of Individually Judged With Study Group Mean Difficulty of Items

Group	Round	*M*	*SD*
Heuristic	Practice	.67	.20
Heuristic	Round One	.72	.10
Heuristic	Round Two	.79	.05
Normative	Practice	.77	.15
Normative	Round One	.78	.08
Normative	Round Two	.93	.03

Table 7-37 shows the means and the standard deviations of the correlations of participants' judged item difficulties with the PLD cut score derived item difficulty estimates for the barely proficient. As shown in Table 7-37, there was a trend of increasing average and decreasing standard deviations of correlations over rounds for both groups. In the practice round of judgment on average, the Normative training group item difficulty estimates correlated higher with the PLD cut score derived item difficulties than the Heuristic training group. The Levene's test of equal variance showed insignificant difference of the variation of the correlations of the Heuristic and the Normative training participants judged item difficulties with the bootstrap PLD cut score *F (1, 20) = .01, p=.92.* The mean difference between the two groups was statistically insignificant at α *=.1,* $|t|$ *(20) =.45, p* =.66.

In the first feedback round of judgment on average, the Normative training group item difficulty estimates correlated higher with the PLD cut score derived item difficulty estimates than the Heuristic training group. The Levene's test of equal variance showed significant difference of the variation of the correlations of the Heuristic and the Normative training participants judged item difficulties with the bootstrap PLD cut score at α *=.1, F (1, 14.09) = 4.13, p=.06.* The mean difference between the two groups was statistically insignificant at α *=.1,* $|t|$ *(14.09) =1.27, p* =.20. In the second feedback round judgment, that followed the construct map feedback for the Normative training participants, the Levene's test of equal variance showed significant difference of the variation of the correlations of the Heuristic and the Normative training participants judged item difficulties with the bootstrap PLD cut score, *F (1, 19.99) = 7.09, p=.02.* On the average the Normative training group's item difficulty estimates correlated significantly higher, with the PLD cut score derived item difficulty estimates than the Heuristic training group's, at α *=.1,* $|t|$ *(19.99) =7.11, p=<.001.*

Table 7-37: Means and Standard Deviations of the Correlations of Judged With Estimated Difficulty of Items at the Proficient PLD Bootstrap Cut Score

Group	**Round**	***M***	***SD***
Heuristic	Practice	.51	.23
Heuristic	Round One	.62	.15
Heuristic	Round Two	.70	.08
Normative	Practice	.55	.19
Normative	Round One	.68	.09
Normative	Round Two	.88	.04

7.3.5. Cut Score Derivatives of Judgments

In this section cut scores results derived from participants item difficulty judgments are presented and for the rounds of judgments. The cut scores were derived for the Heuristic and the Normative training groups by the experiential indicator variables of math specialization, experience teaching at the third or fourth grade level, and teaching, and the overall for the training groups. The independent sample *t*-tests were conducted at $\alpha = .1$. The estimated cut scores for the groups was the average of the sums of the estimated item difficulties of the participants in the involved training. The cut score is an estimate of mean on the true score scale and scale ranges from 0 to number of items on the test (in this case 15, since there are 15 items on the test).

The evidence in this section is presented in the order specified as follows. First, summary statistics tables and discussion is held around cut scores computed by the experiential indicator variables of math specialization, experience teaching at the third or fourth grade level, and teaching along with independent sample *t*-tests of significance results. Second, the table summaries of cut scores by training group results are presented and discussed along with the independent sample *t*-tests of significance of the difference of the means for the groups. Third, summary statistics results of discrepancies of the training groups cut scores from their group mean and the bootstrap PLD cut score criterion are presented and discussed.

Table 7-38 shows the results of independent sample *t*-tests obtained comparing the cut score estimates of math specialists versus non-specialists within the Heuristic and the Normative training groups. As shown in Table 7-38, in the practice round of judgment on average, the Heuristic training group's non math specialists estimated a significantly higher cut score than the Heuristic training group's math specialists The cut score estimated by the non-math specialists in the practice round was also statistically significantly higher and different from the PLD bootstrap cut score estimate. However in the feedback rounds of judgment their appeared alternating pattern of relative magnitude of cut scores estimated by the math specialists and non-specialists with the specialists estimating higher cut score in the first feedback round and the non-specialists in the second feedback round albeit the mean difference for the groups was statistically insignificant. The average cut score for the Heuristic training math specialists and non-specialists was also significantly higher than the bootstrap PLD cut score estimate in the feedback rounds.

On the other hand, in the practice round of judgment on average, the Normative training group math specialist's and non-specialist's cut score estimates were identical and were statistically insignificantly different from the bootstrap PLD cut score estimate. In the feedback rounds of judgment on average the non-math specialists cut scores were higher than those of the math specialists. The difference in the average cut scores of the Normative training group specialists and non-specialists approached significance in the first feedback round and became statistically significant in the second round of judgment that followed construct map feedback. However both the math specialists and non-specialists cut scores were significantly higher than the PLD cut score estimate in the feedback rounds of judgment.

Table 7-39 shows the results of independent sample *t*-tests obtained comparing the cut score estimates of participants with experience teaching at the third or fourth grade level with those of participants that had not taught at these grade levels. In the practice round of judgment on average, the Heuristic training group participants with experience teaching at the third or fourth grade level estimated a higher cut score than the Heuristic training group’s participants without experience teaching at the third or fourth grade level The mean difference between the two groups was statistically insignificant and the cut score estimate of the participants that had taught at the grade levels was also slightly significantly higher than the bootstrap PLD cut score estimate. In the feedback rounds of judgment on average the Heuristic training group participants without experience teaching at third or fourth grade level estimated a higher cut score than the Heuristic training group participants with experience teaching at the third or fourth grade level. The mean difference between the two groups was statistically insignificant.

Table 7-38: Independent Sample *t*-Test Comparisons of Cut Scores of Math Specialists Versus Non-Specialists Within Study Groups

Study Group	Round	Math Specialty	*N*	Cut (*M*)	Cut (*SD*)	$\lvert t \rvert$ [a]	*p*-value
Heuristic	Practice	Yes	6	7.93	1.78	2.87	0.02
		No	4	10.69	0.77		
Normative	Practice	Yes	8	8.47	1.08	0.00	1.00
		No	3	8.47	0.29		
Heuristic	Round One	Yes	6	10.08	1.31	0.24	0.82
		No	4	9.89	1.15		
Normative	Round One	Yes	8	10.25	0.75	1.56	0.15
		No	3	11.13	1.11		
Heuristic	Round Two	Yes	6	9.97	0.93	0.37	0.72
		No	4	10.18	0.77		
Normative	Round Two	Yes	8	10.21	0.54	2.14	0.06
		No	3	11.18	1.01		

[a] *Note: t-* distribution is $\lvert t \rvert$ *(8) f*or the Heuristic and $\lvert t \rvert$ *(9)* for the Normative training group

As shown in Table 7-39, in the practice round of judgment on average, the Normative training participants with experience teaching at the third or fourth grade level cut score estimates were higher than that of those without experience teaching at the third or fourth grade levels. However the mean difference between the two groups was statistically insignificant and statistically insignificantly different from the bootstrap PLD cut score estimate. In the feedback rounds of judgment the Normative participants with experience teaching at the third or fourth grade cut score estimates on average were higher than the Normative training group’s without experience teaching at the third or fourth grade level. The mean differences between the two groups was statistically insignificantly different from each other but were significantly higher than the bootstrap PLD cut score estimates.

Table 7-40 shows results of independent sample *t*-tests obtained comparing the cut score estimates of participants with teaching experience with that of those participants without teaching experience. Across the practice and feedback rounds of judgment the Heuristic training group participants without experience teaching consistently estimated a higher cut score than the Heuristic training group's participants with experience teaching. The mean difference between the cut score of the two groups was statistically insignificantly different in all the rounds of judgment. However, while the Heuristic training group participants without teaching experience average cut score was considerably significantly higher than the bootstrap PLD cut score estimate in the feedback rounds, that of those with teaching experience was slightly significantly higher across the rounds of judgment.

Table 7-39: Independent Sample *t*-Test Comparisons of Cut Scores of With Versus Without Teaching Experience at the Third/Fourth Grade Within Study Groups

Study Group	**Round**	**Taught Grade(3 or 4)**	***N***	**Cut (*M*)**	**Cut (*SD*)**	$\lvert t \rvert$ [a]	***p*-value**
Heuristic	Practice	Yes	5	9.83	1.65	1.31	.23
		No	5	8.24	2.16		
Normative	Practice	Yes	3	8.67	0.50	0.50	0.63
		No	9	8.37	0.98		
Heuristic	Round One	Yes	5	9.78	1.23	.58	.58
		No	5	10.23	1.14		
Normative	Round One	Yes	3	10.67	1.38	0.50	0.63
		No	9	10.37	0.74		
Heuristic	Round Two	Yes	5	10.04	.67	.04	.97
		No	5	10.06	1.05		
Normative	Round Two	Yes	3	10.50	1.43	.28	.79
		No	9	10.34	0.63		

[a] *Note: t*- distribution is $\lvert t \rvert$ *(8) f*or the Heuristic and $\lvert t \rvert$ *(10)* for the Normative training group

In Table 7-40, consistently the cut scores of the Normative training participants with experience teaching cut score estimates were higher than that of those without experience teaching. The mean difference between the two groups was statistically insignificantly different across the rounds of judgment. Both groups cut score estimates were statistically significantly higher than the bootstrap PLD cut score estimate in the feedback rounds of judgment.

Table 7-40: Independent Sample *t*-Test Comparisons of Cut Scores of With Versus Without Teaching Experience Within Study Groups

Study Group	Round	Taught	*N*	Cut (*M*)	Cut (*SD*)	$\lvert t \rvert$ [a]	*p*-value
Heuristic	Practice	Yes	8	8.77	2.11	0.81	0.44
		No	2	10.08	1.45		
Normative	Practice	Yes	9	8.46	0.98	0.13	0.90
		No	3	8.38	0.60		
Heuristic	Round One	Yes	8	9.86	1.24	0.78	0.46
		No	2	10.60	0.99		
Normative	Round One	Yes	9	10.51	0.88	0.40	0.70
		No	3	10.27	1.01		
Heuristic	Round Two	Yes	8	9.88	0.81	1.35	0.21
		No	2	10.73	0.68		
Normative	Round Two	Yes	9	10.42	0.83	0.30	0.77
		No	3	10.25	0.93		

[a] *Note: t*- distribution is $\lvert t \rvert$ *(8) f*or the Heuristic and $\lvert t \rvert$ *(10)* for the Normative training group

The cut scores were estimated by training study group. The overall cut score estimated by the Heuristic and Normative training groups across the rounds of judgments were obtained and the results are presented as follows. As shown in Table 7-41, the Normative training group's estimated cut score in the practice round of judgment on average was lower (8.44) than that for the Heuristic training group (9.03). However in the feedback rounds one and two, the Normative training group cut scores on average were higher (10.45 and 10.38, respectively) than those of the Heuristic training group (10.01 and 10.05, respectively). Meanwhile, the practice round cut score estimates of both groups related most with the PLD estimate and was enclosed by the confidence interval. On the other hand, the cut scores estimated by both groups in the feedback rounds of judgment were higher than that of the practice round and were not enclosed by the confidence interval (see Table 7-32 for these comparisons). The Heuristic training group's cut score increased across the rounds of judgment. The Heuristic training group's cut score increased slightly by .04 points between feedback rounds one and two while the Normative training group's cut score decreased by .07 points in round two that followed the construct map feedback but was still higher than that of the Heuristic training group. In contrast, the standard deviation of cut scores consistently decreased across rounds of judgment for both groups.

Independent sample *t*-tests were conducted to investigate the significance of the observed differences between the cut scores recommendations of the Heuristic and the Normative training group across the rounds of judgments. The results of the *t*-tests are discussed and the table summary of cut score statistics across the rounds of judgment are presented as follows. In the practice round of judgment on average, the Heuristic training group Proficient cut score estimate was higher than the Normative training group. The mean difference between the two groups was statistically insignificant at α *=.1*, $|t|$ *(20) =.93, p* =.36. Both groups cut scores were not significantly different from the bootstrap PLD cut score estimate. In the first feedback round of judgment on average, the Normative training group's cut score estimate was higher than the Heuristic training group's. The mean difference between the two groups was statistically insignificant at α *=.1*, $|t|$ *(20) =1.01, p* =.33. In the second feedback round of judgment that followed the construct map feedback, on average the Normative training group's cut score estimates was higher, than the Heuristic training group. The mean difference between the two groups was statistically insignificant at α *=.1*, $|t|$ *(20) = .94, p*=.36.

Table 7-41: Cut Score Estimates Statistics by Study Groups

Study Group	**Sample Size(*N*)**	**Round**	**Estimated Cut Score**	**Cut Score Standard Deviation**
Heuristic	10	Practice	9.03	1.99
Heuristic	10	Round One	10.01	1.19
Heuristic	10	Round Two	10.05	0.83
Normative	12	Practice	8.44	0.87
Normative	12	Round One	10.45	0.87
Normative	12	Round Two	10.38	0.82

Also computed were the summary statistics of deviations of participant's cut scores from their training group mean and the PLD cut Score criterion. Table 7-42 shows the summary statistics of discrepancies of participants cut scores from their training group mean cut score. From Table 7-42, one can observe that the discrepancies from the group mean cut score for both the Heuristic and the Normative training group decreases across feedback rounds and gets closer to 0 showing convergence of recommendations to the training group mean and that participant's recommendations are unbiased estimates of the training group mean. The standard deviation for the discrepancies also decreases across rounds showing also greater convergence to the group mean.

Table 7-42: Deviation of Participants Cut Score from the Training Group Mean Cut score.

Study Group	**Sample Size(*N*)**	**Round**	**Average of Deviations**	**Standard Deviation of Deviations**
Heuristic	10	Practice	.004	1.99
Heuristic	10	Round One	-.005	1.19
Heuristic	10	Round Two	.001	.83
Normative	12	Practice	.003	.87
Normative	12	Round One	-.003	.87
Normative	12	Round Two	.002	.82

Table 7-43 shows summary statistics of discrepancies of participants cut scores from the bootstrap estimated PLD cut score. From Table 7-43, it can be observed that for the Heuristic training group the mean discrepancies from the PLD bootstrap cut score estimate increased over the rounds and all in the positive direction. For the Normative training group, the mean discrepancy increased from practice to round one but decreased from round one to round two following the construct map feedback. The standard deviations of discrepancies from the PLD consistently went down across rounds for both the Normative and Heuristic training groups.

Table 7-43: Deviation of Participants Cut Score from the PLD Cut Score

Study Group	**Sample Size(*N*)**	**Round**	**Average of Deviations**	**Standard Deviation of Deviations**
Heuristic	10	Practice	.60	1.99
Heuristic	10	Round One	1.86	1.19
Heuristic	10	Round Two	1.90	.83
Normative	12	Practice	.01	.87
Normative	12	Round One	2.30	.87
Normative	12	Round Two	2.23	.82

Chapter Eight: Discussion

The Angoff method was critiqued for been too cognitively complex for participants to perform in the context of its use with the National Educational Assessment Program (NEAP) as to undermine its usefulness for practical purposes (Shepard et al., 1993). Attempts to address cognitive complexity of the Angoff method tasks through training so far, have either introduced instruction for the first component task of conceptualizing the target group of students or feedback procedural modification and without substantive consideration of the knowledge and skills requirements of the tasks. Other attempts to address this cognitive complexity have introduced alternative methods that do not require participants to generate conditional probabilities and also without substantive considerations and especially of the tradeoff between cognitive complexity of a task and accuracy of outcomes.

This dissertation set out to address the cognitive complexity of the Angoff method tasks instead through designing and testing the effectiveness of a criterion-referenced training program with instruction, practice, and feedback tailored to its knowledge and skills requirements. It drew from diverse knowledge bases, the crux of which was the cognitive psychology probability judgment heuristics literature for cognitive task analysis (CTA) and for designing the training. The empirical study applied the conceptual and methodological products of CTA to evaluating the effectiveness of the Heuristic versus the Normative training programs and, the training operations of instruction and practice versus feedback. In the section that immediately follows, the theoretical, conceptual, and methodological intellectual merits of this dissertation are elaborated upon.

8.1. Intellectual Merits of Dissertation

To my knowledge this dissertation is the first to design a training program for a standard setting method task by relying on diverse knowledge bases and in particular the cognitive psychology literature. This CTA approach also served to address the gap in the standard setting literature of lacking theories and frameworks for research. The CTA approach yielded far more insights than would be obtained by mere introspective laboratory reports of the participants and besides was cost effective. It not only illuminated who needs to be trained and what to train, it also illuminated how to design and evaluate training. Because this CTA approach paid off, it is deemed necessary to devote this section to the description of how the task of CTA and design of the training was approached.

Summarily, the CTA and training design research effort was a product of deductive reasoning. It was approached by drawing from the diverse knowledge bases that study the concepts and processes involved in the tasks of the Angoff method. The conceptual and methodological contributions to Angoff method research literature that emanated from reliance on these knowledge bases includes: (1) the heuristic process curriculum (2) the heuristic cognitive process model; (3) the heuristic training program; (4) the comprehensive cognitive and non-cognitive judgment factors conceptual framework; and, (5) comprehensive substantive and technical training evaluation framework.

The subsections that follow elaborate on how these theoretical, conceptual, and methodological products came about and in the logical order as stated above. My hope is that sufficient description of the process used in designing the Angoff method training program would provide guidance and stimulate subsequent sustained efforts to design criterion-referenced training for other standard setting method tasks.

8.1.1. Theoretical Framework for the Angoff Method

The first task of CTA was to understand the philosophical views of measurement science and underlying conduct of research. Two philosophical views were identified namely: the realist and the operationalist view (Dingle, 1950; Hand 1996; Mari, 2005; Michell, 1990, 1999). Because of the goal of this dissertation of veridicality the preferred view was the realist, according to which the role of measurement science is to discover something about reality. The measurement literature in addition suggested what to measure, how to measure, how to evaluate measures, and the scales of measurement (Raykov & Marcoulides, 2011; Steven, 1946).

Following the understanding of the philosophical views about measurement, because the variable of interest in the Angoff method task is test item difficulties, however measured on the probability metric, the next step was to understand the interpretations of probability. The statistics literature identified two schools of thought of probability namely: the objective and subjective probabilities (Popper, 1959). The objective probability, that is the propensity and the frequentist schools of thought were preferred in this dissertation because they fit well with the dissertation research goal of veridicality. The objective probability school of thought helped to clarify who should be trained, how to augment for deficiencies in knowledge and skills through training design, and how to evaluate the training.

Next, because the Angoff method was identified as a judgment task in the standard setting research literature, it was necessary to understand the concept and processes of judgment. Cognitive psychology theories formed the crux to understanding of the notion of judgment specifically, about how people make judgments and as it pertain to the concept of probability. The cognitive psychology probability judgment literature was vested with understanding the processes underlying subjective probability judgment (Beach & Braun, 1994; Gigerenzer, 1994).

However, the objective view of probability was preferred because of this dissertation's goal of fostering veridicality of standard setting outcomes.

The judgment literature identified three theories of how people perform the task of judging probabilities namely: the normative, the prescriptive, and the descriptive theories (Baron, 2000). The normative theories are idealized ways of performing the probability judgment task and are descriptive of how super intelligent or people with advanced statistical knowledge perform the task. The descriptive theories are intuitive strategies that are descriptive of how the average person without advanced training in statistics performs the task of probability judgment. The prescriptive theories which this dissertation joins, recommend approaches to performing the probability judgment task to approximate the normative ideals. This dissertation drew from the descriptive cognitive theories for the task of prescribing training for the Angoff method tasks.

The basis for the descriptive theories of conditional probability judgment task was the notions of bounded and unbounded human rationality which distinction originated from Simon (1957). The unbounded view of rationality assumes no real world and internal constraints to human judgment so that in accordance with this perspective humans have limitless capacity to process all available information in their judgment deliberation. In contrast, the bounded perspective accounts for real world constraints such as limited time, resources, incompleteness of information, and processing limitations in human judgment performance.

Two research programs in the probability judgment literature based on the bounded notion of human rationality in historical order of ascendancy: the heuristics and biases (Tversky & Kahneman, 1974) and fast and frugal heuristic programs (Evans, 2008; Gigerenzer et al., 1999; Gigerenzer & Gaissmaier, 2011) indicated the capabilities of humans to make probability judgments. These two areas of research corroborated cognitive complexity of the task of judging conditional probabilities for the average person and suggested heuristic principles that are quite generalizable to different task situations. The major input from the heuristics and biases program of research for the training design were two heuristic principles that underlie most probability judgment deliberations namely: the representativeness and the availability heuristics (Tversky & Kahneman, 1974). The fast and frugal heuristic perspective (Gigerenzer et al., 1999) in addition, suggested building computational models of heuristics and broadened the model of judgment error.

Going beyond the representativeness and the availability heuristics labels, the fundamental cognitive processes of categorization and recall were identified as underlying the representativeness and the availability heuristics, respectively. It was deduced by extension that the fundamental cognitive processes of categorization and recall underlies the Angoff method probability judgment task. Therefore it was also important to understand the theories of memory and categorization. The inputs from the reviewed memory research for designing the training were that working memory and long-term memory systems are most probably engaged in Angoff method tasks and should be considered for recruitment purposes (Atkinson & Shiffrin, 1968; Baddeley & Hitch, 1974; Murdoch, 1962). The Angoff task necessarily involves free recall of information relevant to the judgment task (Staresina & Davachi, 2006). The levels of processing principles (Craik & Lockhart, 1972) were useful guides for design of training practice activity. Imagining future events involves more constructive processes than retrieving past events, so that retrieving past events should be emphasized in instructions (Addis, Wong, & Schacter, 2007). Selective recall of information introduces errors in judgments, so that participants should be encouraged to engage in extensive recall process (Tversky & Kahneman , 1974). The episodic memory becomes constructive over time (Brown & Kulik, 1977). Therefore, participants should be recruited with recent experience with student population.

It was also important to understand the theories of categorization. Three approaches to categorizing objects that rely on similarity judgment were identified namely: definition, prototype, and exemplar (Murphy, 2002; Sloman & Rips, 1998). The probabilistic approaches namely, the prototype and the exemplar were considered the more relevant for the Angoff method. The rationale for preference of the prototype and the exemplar was that they fit the stochastic framing of the Angoff method tasks and of test taking process as generator of chance events. Two theories were identified about similarity judgment process namely: the geometric and feature matching theory of similarity (Tversky, 1977).

The feature matching theory of similarity judgment was the preferred and the adopted prescriptive model of similarity judgment for the training. The feature matching theory highlighted the need to identify the respects with which similarity judgments are made. Therefore, because the Angoff method task involves estimating item difficulties it was essential to understand the features with respect to which test items could be categorized. The conceptual framework of the knowledge and skills measured by the tests suggested essential features for item difficulty judgment and was provided by the Michigan Curriculum framework (MCF) and the Webb’s Depth of Knowledge (DOK) levels (http://www.michigan.gov/mde).

Ensuing understanding of judgment factors and processes, the last step of CTA was to understand the potential impact of social interaction on judgment outcomes. Because of the social interaction involved in the prevailing Angoff standard setting feedback process, it was necessary to understand the potential impact of discussion on judgment performance. The social psychology literature as reported by Fitzpatrick (1989) identified the potential ills of discussion.

Specifically, the group discussion polarization phenomenon was suggested as the major potential deleterious impact of discussion which was directly relevant to design of test stimulus and the training feedback. It was considered appropriate social influence cognitive learning through the exchange of information. This concludes discussion of the research sojourn into the diverse knowledge bases, theoretical decisions made, and the rationales. The rest of the subsections on the intellectual merits of the dissertation that follow elaborate on the conceptual and methodological products of the CTA.

8.1.2. The Heuristic Training Curriculum and the Cognitive Process Model

The primary products of CTA based on reviewed theoretical frameworks for the Angoff method were the curriculum and cognitive process model that informed instruction, practice, and feedback design. The curriculum specified the knowledge requirements while the cognitive process model specified how to perform the Angoff method tasks. In McGinty's (2005) language the "factors" deduced underlying the Angoff method tasks based on the feature matching similarity theory of categorization were categorical content domain knowledge and skills constructs. The "Black box" of the Angoff method tasks deduced from the representative and the availability heuristic principles was the interaction of categorization and recall. The scales of measurement theory (Steven, 1946), in addition identified ranking as a fundamental knowledge and skill requirement underlying all measurement operations and therefore as a pre-requisite skill set for the Angoff method task. Hence, the cognitive process model was formulated in terms of interaction of categorization and recall.

These heuristic principles were formalized in the training, which instruction placed constraints on the features considered in the item difficulty judgment task to be categorical knowledge and skills constructs measured by the test while recall was unconstrained. The idea of constraining features was for the participants to pay optimal attention to knowledge and skills measured by the test items while recall was left unconstrained so that participants would engage in extended and unbiased recall.

8.1.3. The Heuristic Training Program

The heuristic principles of categorization and recall were formalized in the Heuristic training program instruction on the knowledge and skills measured by the test and for the Angoff method tasks for the idealized situation when the participants have firsthand experience with the student and item population. The heuristic training instruction on the knowledge and skills measured by the test incorporated DOK instruction. The DOK is considered as the substantive domain item difficulty ordering which ordering might differ from the empirical item difficulty ordering (see, Hartel & Lorie, 2004, for the notion of substantive domain item difficulty ordering). The training instruction for the Angoff method in this implementation prescribed thinking about real group of students that fit the performance category, categorizing items in terms of their knowledge and skills features that include DOK and content strands, recalling and using similar experienced item instances as reference class for their judgment. The rationale for the instruction to think in terms of real students was because of the goal of veridicality and because of the research evidence in the standard setting literature that suggests that teachers are better able to think in terms of real examples of categories.

It was conceived that the prevailing practice activity of taking the test might foster instead processing of surface features of test items instead of more conceptually oriented features. Therefore the Heuristic training practice activity was designed to include the principles of categorization and rank ordering. The practice task of categorization entailed designating the items to content domain knowledge and skills that include the DOK and content strands. The practice task of rank ordering entailed ranking items in terms of difficulty and was informed by Steven's (1946) levels of measurement theory. These practice activities were meant to reinforce conceptual understanding of the content domain knowledge and skills measured by the test and the substantive domain item difficulty ordering of the test items.

The feedback design as conceived ideally should be based on reinforcing through instruction and practice activities knowledge and skills constructs measured by the test items, substantive domain item difficulty ordering of items, the performance category, and the heuristic judgment strategy. However, for operationalization in this dissertation was mixed to include the prevailing discussion approach on final outcomes of judgments of the participants, empirical data, as well as the intermediate outcome discussions on the participants' designations of item DOK.

Meanwhile, the training for the comparison group, called the Normative training was meant to simulate the prevailing Angoff method training. For instance, the Normative instruction on the knowledge and skills measured by the test indicated the categories of topics covered but did not specify the complexity of the test items. The Angoff method instruction was the hypothetical and group formulation. The practice activity was the business as usual taking the test, however included rank ordering items in terms of difficulty. Although it is not typical to ask participants to rank order items in terms of difficulty, it was incorporated for the Normative training to facilitate comparability of the two training. The feedback simulated the final outcome types and included detailed empirical data feedback types as well as discussions around the participant's judgment outcomes and the feedback norms.

8.1.4. The Conceptual Framework of Factors Impacting Probability Judgment

CTA also yielded the comprehensive conceptual framework of possible sources of errors in judgment, as deduced based on statistical theories and the fast and frugal heuristic perspective, for designing and testing Angoff method training effectiveness. The cognitive psychology training and the fast and frugal heuristics literature identified non-cognitive constructs namely: motivation, engagement, and emotion as extraneous variables that could potentially confound with training effectiveness. Other potential confounds identified includes background characteristics such as teaching experience and stimulus characteristics. The utility of the comprehensive conceptual framework of potential extraneous sources of judgment errors is that it can be applied directly in planning, designing, and evaluating Angoff method training programs. For instance, the framework suggests the variables to measure and test for balance in an ANOVA and experimental design framework to the study of training method effectiveness.

8.1.5. Kirkpatrick's Training Evaluation Framework

So far, standard setting training evaluation have focused on the reaction and transfer levels while the knowledge and skills acquisition are neglected. Transfer evaluation also focuses on the technical quality of outcomes while overlooking substantive meaningfulness of judgment process and outcomes. This state of affairs is due to the assumption that participants understand the tasks and are able to do it. This practice also highlights the identified lacking substantive cognitive theory guiding standard setting methods and training, a gap addressed by this study. This dissertation demonstrated application of a comprehensive training evaluation framework based on Kirkpatrick's criteria measures. The Kirkpatrick's framework identifies five criterion measures for evaluating a training program namely: (1) reaction or satisfaction of participants (2) knowledge and skills acquisition; (3) transfer of learning to tasks; and, (4) costs and benefits. In addition to evaluating the final outcomes of the probability judgments and outcomes, also obtained were measures of the participant's performance on the pre-requisite tasks of categorization, recall, and rank ordering of items in terms of difficulties. The utility of this comprehensive training evaluation approach when applied increasingly in evaluating training programs is that it would substantially extend the knowledge base through providing reliable and valid information about what training intervention works, why, and how.

8.1.6. Multidimensional Scaling Training Evaluation Framework

So far, standard setting research evaluation focuses on the outcomes while neglecting the judgment processes, consequently little is known about the factors and processes that underlie standard setting participant's judgments (McGinty, 2005). Evaluation efforts have also emphasized correlational indices and item response theory models (e.g., Englewood & Stone, 1994; Reckase, 2006; Wyse, 2009). This dissertation extended judgment validity evaluation framework by adopting and demonstrating application of both correlational indices and the multidimensional scaling distance metric based analytic framework to evaluating informed process and substantive meaningfulness of judgment outcomes in relation to the heuristic process model assumptions.

8.1.7. Bootstrap Resampling PLD Cut Score Estimation Method

This dissertation is also the first to apply the bootstrap resampling approach to operationalizing the PLD for the purpose of cross validating Angoff method judgment outcomes. The bootstrap resampling estimates were also based on criterion-referencing to the knowledge and skills measured by the test items. The approach to estimating cut score by the bootstrap method was based on the same underlying principles of categorizing items by their content domain knowledge and skills features as prescribed in the training instructions. This concludes discussion of the conceptual and methodological contributions of this dissertation. The immediately following paragraph presents the finding of the empirical study.

8.2. Empirical Study Findings

There was no observed statistically significant difference between the Heuristic and the Normative training on the measured extraneous variables that could potentially confound with training effectiveness. Therefore, the findings discussed in this section comprise the aspects of evidence from data analysis that addressed the hypotheses that the Heuristic training would enhance pre-requisite knowledge and skills acquisition and improve item difficulty judgment outcomes. The evidence is presented in accordance to Kirkpatrick's framework levels of measures for evaluating training effectiveness. The evidence in Kirkpatrick's category of knowledge and skill acquisition and transfer are presented together for convenience. The indicators of performance on pre-requisite tasks pertain to knowledge and skills acquisition while indicators of performance on the Angoff method tasks pertain to transfer. The discussion of findings in this section also highlights the variables, indices, and analytic methods that generated the evidence.

8.2.1. Training Effectiveness: Reactions

The evidence based on self-reports of the participants indicated as expected, that the Heuristic training group on average expressed statistically significant higher satisfaction with the training intervention and with the performance of the facilitators. There was also insignificant difference of the variation of the responses of the Heuristic and the Normative training participants to the questions that elicited their satisfaction with the training intervention and the performance of the facilitators. Also, the Heuristic training participants raised fewer issues about the adequacy, appropriateness, usefulness, or organization of the training. The responses of both the Heuristic and the Normative training participants to open ended questions that elicited what they found useful about the training showed consensual value of group discussions and the Angoff method. Besides, the responses to open ended items indicated enthusiasm to know more about other standard setting methods that are used in practical settings.

8.2.2. Training Effectiveness: Knowledge and Skills and Transfer to Tasks

The findings discussed in this section pertain to evidence of the Heuristic cognitive process, pre-requisite knowledge and skills acquisition and, transfer to the Angoff method tasks. Evidence presented in this section and, in order includes: self-reports of the participants about their task performance, actual participants' performance on the pre-requisite knowledge and skills tasks, and transfer to the Angoff method item difficulty judgment tasks. The evidence on transfer of knowledge and skills to Angoff method tasks are presented separately for the practice and instruction, and the feedback rounds of judgment. The rationale for presenting evidence about the performance of the participants after instructional and practice activities versus feedback separately is to facilitate addressing the research questions and to effectively draw the contrasts between the impacts of these training operations on the participant judgment performance.

A. Self-Reports

The findings discussed in this part pertain to four dimensions of training effectiveness indicators elicited by Likert scale evaluation survey. The dimensions and, as presented in order in the paragraphs that follow were (1) impact of five potential predictors of item difficulty in the judgment deliberations of the participants namely: feedback and discussion, educational and classroom experience, knowledge and skills constructs measured by the items, item quality, and the PLD; (2) understanding of instruction, tasks and feedback; (3) confidence in conceptions and recommendations; and, (4) rationales for responsiveness to feedback.

(1) Judgment factors - the Heuristic training participants expressed higher consideration of the elicited five factors in their judgment deliberations than the Normative training participants. The significant findings based on the Likert scale responses of the participants and which conformed to expectations were that the Heuristic training participants expressed statistically significant higher weighting of their educational and classroom experiences, while the Normative training participants expressed statistically significant higher weighting of feedback and discussion in their judgment deliberation. There was also statistically significant higher variation of the responses of the Normative training participants than those of the Heuristic training participants on the impact of their classroom and educational experience in their judgments. Additionally, other equally important although statistically insignificant findings were that the Heuristic training participants expressed higher impact of the constructs measured by the test and the item quality, while both groups expressed equal impact of PLD in their judgment deliberation.

(2) Understanding - overall the Heuristic training participants expressed higher understanding of instructions, tasks, and feedback. Although the difference in terms of average and variation of responses between the two groups was not statistically significant, however they were nearing significance.

(3) Confidence - the Heuristic training participants expressed higher although statistically insignificant confidence than the Normative training participants both in their conception of the barely proficient students and in their cut score recommendations.

(4) Rationale for responsiveness to feedback – the Heuristic and the Normative training participants' responses indicated overall responsiveness to feedback although there was less clarity as to whether the influence was due to better learning. For instance, the Heuristic training participants responded more affirmatively to not adjusting their recommendations in the light of feedback as a result of not learning from discussions and for not wanting to use others ideas albeit on the average the difference between the two groups was not statistically significant. On the other hand, the Normative training participants responded more affirmatively to not adjusting their recommendations in the light of feedback due to confidence although on the average the difference between the two groups was also statistically insignificant.

B. Performance on Pre-Requisite Tasks

There are two types of evidence based on the measures of pre-requisite knowledge and skills requirements of the Angoff method task of judging item difficulties in accordance with the heuristic principles. The two types of evidence and, as presented in the order are as follows: (1) recall and recognition tasks performance; and, (2) performance on rank ordering of items in terms of difficulty.

(1) Recall and recognition - there was no observed difference in the performance of the Heuristic and the Normative training participants on the pre-requisite task that measured their ability to recall information and in their recognition of the similarity of the knowledge and skills measured by test items. In terms of specifics, both the Heuristic and the Normative training participants had a recall rate of 99% for items that were replicated on the two tests that were used for the study. Also, both the Heuristic and the Normative

training participants had approximately 80% recognition rate of the similarity of the knowledge and skills measured by the items that appeared on the two tests.

(2) Rank ordering items in terms of difficulty - in the practice round of judgment that followed training instruction and practice activities, the Heuristic training group relative item difficulty judgments on average correlated higher with the empirical item difficulty ranks based on entire fourth grade students responses in 2005, while the Normative training group's correlated higher with their study group mean item difficulty ranks. The correlation of the Heuristic and the Normative training group relative item difficulty ranks with the item difficulty ranks based on empirical data was moderate, approximately .50 for both groups and the mean difference for the groups was statistically insignificant in the practice round of judgment. While the correlation of the Normative group item difficulty ranks with their study group item difficulty ranks in the practice round on average was approximately .70 that of the Heuristic group was approximately .60.

On the other hand, both group's judgments of the relative item difficulty ranks on average correlated higher with the item difficulty ranks based on the aforementioned two feedback norms while the standard deviations of the correlations consistently declined in the feedback rounds of judgment. Also, the Normative training group's rank ordering of items correlated considerably higher with the ranks of the items based on these two feedback norms in the feedback rounds of judgments. The difference in average correlation of the two group's judgments of the relative item difficulties with these two feedback norms tended towards and became significant in the second feedback round. Specifically, the Normative training participants judged relative item difficulty ranks on the average correlated significantly higher (in the .90's) with the item difficulty ranks based on the feedback norms in the second feedback round that followed construct map feedback.

C. Performance on the Angoff Task After Instruction and Practice

The findings presented in this part pertain to the results of analysis of the practice round of item difficulty judgments that followed instructions and practice activities for the two groups. This practice round of judgment was supposed to provide evidence about the cumulative impact of the qualitatively different instructional and practice activities for the Heuristic and the Normative training groups. The evidence presented pertain to result of check for extent to which the participants translated the Proficient performance standard well and for how well their judgment data conform to expectations about the knowledge and skills characteristics of the PLD, the test and about the processes underlying their judgments. The findings are broken up further into five logically ordered evidential types as follows: (1) unidimensionality of the knowledge and skills constructs measured by test items and unimodality of the Proficient student performance category; (2) item substantive domain construct coherence; (3) predictability of participants judgments by experiential and background factors; (4) substantive meaningfulness and technical quality of item difficulty judgments and, (5) substantive meaningfulness and technical quality of cut score judgments.

(1) Unidimensionality of the knowledge and skills constructs measured by the test items and unimodality of the Proficient student performance category - The first step taken in the evaluation of the participant's practice round judgment performance was to check their judgments for how well they recover holistic assumptions made about the nature of the knowledge and skills of the students as delineated in the Proficient PLD and measured by the test items. Principal coordinates analysis exploration of the practice round judgment

data suggested that the assumption that a single dimension of knowledge and skills construct is measured by the test and a single student performance category underlies the Proficient performance standard were tenable. In technical terms the assumption that a single dimension of knowledge and skills construct is measured by the tests is that of unidimensionality while that of a single student performance category is unimodality. Therefore, it was considered appropriate to describe the practice round judgment data for both the Heuristic and the Normative training group with a single dimension of item difficulties and a single cut score.

(2) Check of item substantive domain construct coherence - Because the judgment data fit the assumptions of unidimensionality and unimodality, the next line of action taken was to check for the content domain construct predictors of clustering of items in a two dimensional principal coordinates plot. The findings as expected were that the Depth of Knowledge (DOK) level and content strands of the mathematics test items were predictive of clustering of the items. The Depth of Knowledge (DOK) level of the items was the predominant predictive construct of clustering of items along the first principal coordinate.

(3) Check of predictability of participants judgments by experiential and background factors – Both the Heuristic and the Normative training participants tended to cluster together in the two dimensional principal coordinates plots in terms of known indicators of experience such as whether they had taught or were math specialists in the practice round of judgment.

(4) Substantive meaningfulness and technical quality of item difficulty judgments – The reliability of item difficulty judgments of the participants was evaluated based on

correlating Euclidean distance matrices of the judgments on items that were replicated on the Practice and the Real test. The findings were that the Heuristic training group difficulty judgments for the replicated items in the practice round remained more consistent with their judgments in the feedback rounds than those of the Normative group.

Meanwhile, construct validity of the Heuristic and the Normative training group judgment were evaluated based on correlating them with four other criterion measures of the item difficulty. The four criterion measures used to evaluate participants item difficulty judgments were: (1) the modal DOK designated to the items by content experts that participated in alignment study in 2005; (2) empirical proportions of the entire fourth grade students responding correctly to the items in 2005; (3) their study group mean item difficulty estimates; and, (4) item difficulties estimated at the bootstrap PLD cut score. On average, the Normative training group judgments correlated higher with all four criterion measures of item difficulty albeit the mean difference for the two groups was statistically insignificant.

Both groups item difficulty judgment correlations were in the expected direction, they correlated negatively with the DOK designation of the items and positively with all other criterion measures. Except for their study group mean criterion based correlations, the average correlations for both groups with other criterion measures of item difficulty were moderate and hovered in the range of .40 -.60. The Heuristic and the Normative group item difficulty judgments correlated highest with their study group mean estimates and followed in order by: item difficulties estimated at the bootstrap cut score, empirical item difficulties based on the entire fourth grade student responses in 2005, and least with the modal content expert DOK designations of the items.

(5) Substantive meaningfulness and technical quality of cut scores - Cut scores were computed within the Heuristic and the Normative training by experiential indicator variables of math specialization, experience teaching at third or fourth grade, and experience teaching. The findings were that the non-math specialists in the Heuristic training group estimated significantly higher cut score than math specialists. The cut score estimated by non-math specialists in the Heuristic training group was also statistically significantly higher than the bootstrap PLD cut score estimate. On the other hand, non-math specialists in the Normative training group estimated identical cut score as the math specialists which was statistically insignificantly different from the PLD bootstrap cut score estimate.

The Heuristic and the Normative training participants with teaching experience at the third or fourth grade estimated higher cut score than those without experience although the difference between the two groups was not statistically significant. The Heuristic training group with teaching experience at the third or fourth grade cut score was also slightly significantly higher than the PLD bootstrap cut score estimate. The Heuristic training group with no teaching experience estimated higher cut score than those with teaching experience while the reverse was the case for the Normative training group, however the difference between the two groups was statistically insignificant. With regards to training group average, the Heuristic training cut score was higher than that of the Normative training group however both groups cut score was quite reasonable and comparable to that obtained by operationalizing the PLD using the bootstrap resampling method. The mean difference in cut score estimates of both groups was also statistically insignificantly different.

D. Performance on the Angoff Task After Feedback

The feedback rounds of item difficulty judgment in addition to facilitating evidence about the effectiveness of the Heuristic training versus the Normative training was supposed to provide evidence about the third research question which explicitly asked whether feedback had additional impact on the judgment performance of the participants net the impact of instructional and practice interventions and irrespective of training method. The findings are broken up further into the same five logically ordered evidential types as was done for the evidence pertaining to instructional and practice activities as follows: (1) unidimensionality of the knowledge and skills constructs measured by test items and unimodality of the Proficient student performance category; (2)item substantive domain construct coherence; (3) predictability of participants judgments by experiential and background factors; (4) substantive meaningfulness and technical quality of item difficulty judgments and, (5) substantive meaningfulness and technical quality of cut score judgments.

(1) Unidimensionality of the knowledge and skills constructs measured by the test items and unimodality of the Proficient student performance category - Principal coordinates analysis exploration of the feedback judgment data showed that the assumption that a single dimension of knowledge and skills construct is measured by the test and a single student performance category underlies the Proficient performance category were tenable in the feedback rounds. Therefore, it was considered appropriate to describe the feedback judgment rounds data for both the Heuristic and the Normative group with a single dimension of item difficulties and a single cut score.

(2) Check of item substantive domain construct coherence – The Heuristic training participant’s judgments still showed pattern of items clustering by the DOK constructs albeit the separation between the items of DOK level 1 and 2 declined. There was substantially reduced clarity in the clustering pattern in terms of the DOK constructs measured by the test in the feedback rounds for the Normative training group.

(3) Check of predictability of participants’ judgments by experiential factors - The clustering pattern in terms of experiential factors disappeared over the feedback rounds, especially for the Heuristic training group. The clustering pattern in the feedback rounds showed patterns of relationship of judgments with the researcher induced break ups of the participants (i.e., in terms of table groups).

(4) Substantive meaningfulness and technical quality of item difficulty judgments – With regards to reliability, the Heuristic training group item difficulty judgments of the replicated items in the feedback rounds remained more consistent than those of the Normative group albeit both groups judgment on these items remained appreciably consistent.

The correlations of both groups item difficulty judgment with other measures of item difficulty were in the expected direction, they correlated negatively with the DOK designation of the items and positively with all other criterion measures. The Heuristic training group judgments in the feedback rounds on average correlated higher with the DOK designation of the items than those of the Normative training group, the difference between the two groups tended towards and became statistically significant in the feedback round two. The correlations for the Normative training group judgment with the DOK designation of the items consistently declined across the feedback rounds while those of the Heuristic training group remained fairly stable and increased between feedback rounds one and two.

The average of correlations of both training judgments with the empirical proportions of the fourth grade students responding correctly to the items, their study group mean estimates, and the item difficulties estimated at bootstrap PLD cut score showed increasing trend while the standard deviations of the correlations showed declining trend across the feedback rounds. The Normative training group's judgments in the feedback rounds on average correlated higher with the empirical item difficulties based on fourth grade student responses in 2005, their study group mean, and the item difficulties estimated at the bootstrap PLD cut score. The mean difference between the two group's correlations with the latter criterion measures tended towards and became statistically significant in the feedback round two that followed construct map feedback for the Normative training group. In the second feedback round of judgment that followed the construct map, the Normative group correlation with these criterion measures were considerably higher and in the vicinity of .90's while those of the Heuristic training group were in the range of .60-.70.

(5) Substantive meaningfulness and technical quality of the cut scores – In the feedback rounds of judgment their appeared alternating pattern of relative magnitude of cut score estimated by the non-math specialists versus math specialists in the Heuristic training group, with the specialists estimating higher cut score in the first feedback round and the non-specialists in the second feedback round. The cut score estimated by both specialists and non-specialists in the feedback rounds were also statistically significantly higher than the bootstrap PLD cut score estimate. On the other hand, non-math specialists in the

Normative training group consistently estimated higher cut score than the math specialists in the feedback rounds. The mean difference between the average cut score of the non-specialists and those of the specialists in the Normative training approached and became statistically significantly different in the second feedback round of judgment that followed construct map feedback. Also, both the non-math and math specialists cut scores in the Normative training were statistically significantly different from the bootstrap PLD cut score in the feedback rounds of judgment.

The Heuristic group with no teaching experience at the third or fourth grade estimated higher cut score than those with experience while the reverse was the case for the Normative training group. The Heuristic training group with no teaching experience estimated higher cut score than those with teaching experience while the reverse was the case for the Normative training group, however the difference between the two groups was statistically insignificant and were significantly higher than the bootstrap PLD cut score estimate. In terms of training group average, the Normative training cut score was higher than that of the Heuristic training in the feedback rounds of judgment however, both groups cut score were comparable and statistically significantly higher than that obtained by operationalizing the PLD using the bootstrap resampling method.

The Heuristic training group cut score consistently increased across feedback rounds, while that of the Normative training declined in feedback round two of judgment that followed construct map feedback. Also indices of discrepancies indicated that while the cut score estimates of both training groups increasingly converged to their study group mean across the feedback rounds, on the other hand they increasingly deviated from the PLD cut scores estimates.

8.3. Interpretation of Empirical Study Findings

The interpretation of study findings in this section is broken up into four subsections. The first three subsections present interpretation of findings pertaining to specific levels of the Kirkpatrick's training measures evaluation framework. The evidential levels for which findings are interpreted and in order are as follows: (1) reactions and self-reports of task performance; (2) performance on pre-requisite knowledge and skills tasks; and, (3) transfer of the knowledge and skills to the Angoff method tasks. Evidence about transfer of the knowledge and skills to the Angoff method tasks is discussed separately for the training interventions of instruction and practice, and feedback. The interpretation of evidence about transfer address the research questions about fit of judgments to the heuristic model assumptions, substantive meaningfulness and technical quality of judgment process and outcomes. The last subsection presents summary of interpretations.

8.3.1. Participants Reactions and Self Reports of Task Performance

As anticipated the Heuristic training that received qualitatively better training instructions and practice activities expressed significantly higher satisfaction with the training, higher satisfaction with the performance of facilitators, and raised fewer concerns about the adequacy of procedural implementation of the training. In addition, the Heuristic training group expressed higher understanding of instruction, tasks, feedback, confidence in their conceptions, and recommendations. To a considerable degree the participants of both the Heuristic and the Normative training seemed to have followed instructions and to weight higher what was emphasized in training instructions.

For instance, the Heuristic training group expressed significantly higher impact of their educational and classroom experience in their judgment deliberation, while the Normative training group expressed significantly higher impact of feedback in their judgment deliberation. This finding that the Heuristic training expressed significantly higher impact of their educational and classroom experience conformed to expectation, because their training instruction specified reliance on actual experiences in judgment deliberation. The finding that the Normative training group expressed higher reliance on feedback also makes sense because of the less lucid training instruction for them and was suggestive of that in the absence of appropriate instruction and meaningful practice activities that participants may be more reliant on feedback. This finding of higher expressed feedback reliance was also corroborated by actual analysis of the judgment data which also indicated that feedback did have more undesirable influence on the judgment behavior of the Normative participants.

Although not statistically significant the evidence did also indicate that the Heuristic training participants weighted higher the knowledge and skills constructs measured by test items in their judgment deliberation while both groups expressed equal impact of the PLD in their judgment deliberations. The finding of higher expressed impact of knowledge and skills constructs measured by the test by the Heuristic training participants conformed to expectation because their training instruction explicated conceptualizing items in terms of the knowledge and skills they measure. Overall, self-report of the participants indicated that the training instructions and practice activities mattered as much as feedback in the participant’s judgment deliberation.

8.3.2. Pre-Requisite Knowledge and Skills Acquisition

Going a step further in the task of evaluation, there was no observed difference in the performance of the Heuristic and the Normative training on the pre-requisite knowledge and skills tasks of recall and recognition despite the qualitatively different instructions and practice activities. This finding did not conform to expectation in accordance with the depth of processing theory which suggests that categorization constitutes deeper processing of a stimulus and fosters better recall and recognition task performance (Craik & Lockhart, 1972). The expectation was that the practice activity of categorization constitutes deeper processing of meaning as opposed to taking the test which instead might foster more processing of surface features of test items.

The explanation for this latter finding might be because the Normative training group took the Practice test immediately after the knowledge and skills measured by the test was reviewed while the Heuristic training group performed the task of categorizing items along with the practice round of judgment and, the memory test was administered to both groups immediately after feedback on their practice round judgment. Therefore, besides the fact that there was a good number of math specialists in the Normative training, the study design may have given them an edge over the Heuristic group because the Practice test was presented to them first, so that they had more time to process it and also they had opportunity to rehearse it further as a result of feedback discussions. This explanation is also supported by the classical memory theory Murdoch (1962) distinguishing the long-term from the short-term memory which suggested that recency and primacy in the order of presentation of information matters for recall performance.

The prediction is that with better design of the training that allows for the practice activities of categorization versus taking test in the same positional order, larger sample size of test, longer duration between practice and memory tests, that the practice activity of categorization would prove to be more effective in enhancing better learning of test knowledge and skills than taking test. All the same, the finding of near perfect recall and recognition performance of the two groups indicated the participant’s readiness for performing the tasks, no issues with their memory, and also suggested that they processed the Practice test items very well.

With respect to the pre-requisite task of judging relative item difficulties, the Heuristic training group relative item difficulty rank correlated higher with the empirical item difficulty ranks in the practice round that followed training instruction and practice, although the mean difference between the two groups was statistically insignificant. This finding was expected because the practice activity of designating items to DOK was meant to foster knowledge of substantive domain item difficulty ordering (see, Hartel & Lorie, 2004). Also, the finding was that although both groups relative item difficulty ranks correlated moderately with the relative item difficulty ranks based on feedback norms in the practice round that followed training instruction and practice activities, however they correlated considerably higher in the feedback rounds. These relative item difficulty rank correlations were much higher for the Normative training group in the second round of feedback that followed the construct map feedback and the mean difference between the two groups was statistically significant.

The finding that the Heuristic and the Normative training relative item difficulty rank correlated higher with the empirical and their study group relative item difficulty ranks across the feedback rounds indicated that feedback did have the effect of enhancing monotonicity of the relationships of participant’s judgments with the feedback norms. This evidence suggested that restriction of range of participant’s judgment as offered by Wyse (in press) was not the only explanation of the phenomena of increasing correlation of the participant's item difficulty judgments with these feedback norms.

8.3.3. Transfer of Knowledge and Skills to the Angoff Method Tasks

The last step in the task of evaluation was check for transfer of knowledge and skills to the Angoff method tasks. The findings interpreted in this section comprises of those that addressed the research questions. The interpretations specifically address findings of evaluation checks of the informed process underlying participant's judgment for: (1) consistency with the heuristic principles; (2) recovery of assumptions about the content domain knowledge and skills measured by the PLD and the test items; and, (3) relationship of participant's item difficulty and cut scores with similar process generated measures. The interpretation of finding is presented separately for the practice round of judgment that followed training instructions and practice activities, and for the feedback rounds.

A. Evidence for Transfer After Training Instruction and Practice

The findings from analysis of the practice round of judgment data that followed training instructions and practice activities was that the Normative training group performed just about equally well as the Heuristic training group in terms of all indices used for evaluation. The Normative training participant judgment process also fit the judgment heuristic process model and their cut scores were also quite reasonable relative to substantive considerations.

The first evidence to suggest equal performance was that both training groups' judgment data recovered assumptions that a single knowledge and skill construct and a single student performance category underlies the PLD and is measured by the test items. This finding was suggestive of that both groups were able to translate the text of the Proficient PLD categorically in this practice round of judgment (Reckase, 2009).

The content domain knowledge and skills constructs measured by the test were the predictive constructs of item clustering. Specifically, the DOK and content strands of items appeared to be the major predictive factors with the DOK been the most predominant predictive construct of item clustering. This finding indicated that categorization was involved in the participant's judgment deliberation and that DOK and content strands are psychologically meaningful constructs, as well as the level of categorization of the knowledge and skills constructs measured by the items.

This finding of DOK and content strands as predictive of item difficulty judgments also indicated that the factors driving participant's item difficulty judgment deliberation in this practice round were neither metric nor dimensional as is assumed by correlational parametric evaluation approaches (Englewood & Anderson, 1998; Reckase, 2006; Wyse, 2009). Instead, the participant judgment process in this practice round conformed to the monotonic feature matching theory of similarity judgment (Tversky, 1977). Thus, the representativeness judgment heuristic was descriptive of the participant's judgment behavior (Tversky & Kahneman, 1974).

The second evidence to suggest equal performance was that both groups item difficulty judgments correlated moderately with feedback norms in this practice round, however cut score judgments were quite reasonable based on substantive considerations. Although both training groups item difficulty judgments process conformed to the judgment heuristic principles in this practice round, however they were quite reasonable. This finding lend support to the fast and frugal heuristics program contention Gigerenzer et al. (1999) that heuristics strategies are not the only explanatory construct of why human judgments are biased and that in some circumstances consideration of less information in judgment deliberation might be more effective.

It is interesting that even without elaborate training instructions on how to estimate item difficulties, the Normative groups' item difficulty judgment in this practice round that followed instructional and practice activities conformed to the heuristic principles. This finding that the Normative training group judgment process also conformed to the feature matching similarity judgment principles Tversky (1977), lend credence to the judgment heuristic theory. Because the Normative training instruction simulated the prevailing training instruction and practice on how to estimate item difficulties implies that this finding addressed McGinty's (2005) call of illuminating the "Black box" of uninstructed Angoff method participants. The factors underlying the Angoff method participant item difficulty judgment in this instance were the DOK and content strands and their processes fit the feature matching similarity judgment process.

This finding of substantively meaningful and reasonable judgment outcomes in this practice round that followed training instruction and practice for both the Normative and the Heuristic training contrasted with most contemporary results of Angoff studies that seem to suggest instead that judgments of participants are flawed in the absence of performance data (Clauser et al, 2002; Clauser et al., 2009a; Wyse, 2009; Wyse, in press). It generated the speculation that perhaps the heuristic principles are the intuitive way that participants of previous studies might have been using to perform the Angoff tasks prior to feedback rounds but probably wasn't captured by evaluation methods used by previous research studies. It was also suggestive of that the critique of cognitive complexity of the Angoff method (Shepard et al., 1993) would be discredited with appropriate training and evaluation methods.

However, the findings about the impact of training instruction and practice may not generalize to all Angoff method standard setting contexts because of the atypical although ideal performance enhancing procedures introduced by this dissertation. The atypical procedures may have in addition augmented for knowledge and skills deficiencies and may also explain the observed equity in performance of the Normative and the Heuristic training groups. The prominent atypical procedures include: (1) the test was assembled so that items aligned well with the PLD; (2) participants were assigned to table groups for discussion of the PLD to create balance in experience; and, (3) the Normative training participants were also asked to perform the pre-requisite practice task of rank ordering items in terms of difficulties. Thus, with the experimental design that ensured distribution of experience in the discussion tables and with sufficient alignment of items with the PLD, PLD discussion may have augmented for deficiencies in knowledge of the student population even without elaborate instruction.

The prediction is that if items were selected that do not align well with the knowledge and skills of PLD, there might have been more qualitative and quantitative difference between the Heuristic and the Normative group judgment performance in this practice round that followed training instruction and practice activities. The speculation associated with the latter prediction is that with the experimental arrangement that ensured distribution of experience in table groups for discussion of the PLD, the Normative training participants and the Heuristic training participants with deficient knowledge of the student population may have applied the strategy of matching items to a PLD, recalling, and using discussed information about the matching PLD for judgments. Presumably, information indicating relative frequencies of correct response of the target students to items measuring the PLDs was generated during the PLD discussion. This matching item to PLD strategy most likely generated the Normative training judgment and may have yielded reasonable cut score estimates in this instance because math content is hierarchically structured. The hierarchical structure of math content implies that the content strands are not orthogonal to DOK, so that consideration of only content strand and matching to PLD strategy in judgment deliberation might have made it look as if the Normative training explicitly considered DOK in their judgment deliberation as the Heuristic training was instructed. This speculated matching item to PLD strategy is apparently simpler as it may not even require global conceptualization of the target students and deliberate consideration of the knowledge and skills measured by individual test items to apply which may be why the real versus hypothetical instruction seemed to not make a difference.

The observed equal judgment performance despite the qualitative difference in the Angoff method instruction also gave insight that the effectiveness of the instructions to conceptualize real versus hypothetical students and items may be dependent on the experience represented by the participants. For instance, the real and group formulation of instruction may work better in the circumstance that the participants have firsthand knowledge of the student population while the hypothetical and group formulation might work better in the circumstance of deficient firsthand knowledge about the student and item populations and for content experts. It could also be that content experts might be better able to think in abstract terms so that the hypothetical and group formulation instruction might work better in the circumstance of content expertise. The latter seemed to be the case in this dissertation because the Normative training comprised of a good number of math specialists.

Taken together these alternative explanations for no observed differential impact of instruction and practice activities on the judgment performance of the Heuristic and the Normative training suggest that it may not always be the case that without explicit instruction on how to conceptualize items, and integrate knowledge and skills constructs, and experiential information in the task of judging item difficulty, that the participants would be able to figure them out. Also, it may not always be the case that the same heuristic strategy would be intuitive way of performing the Angoff method task. The heuristic strategy that might work for instruction on how to perform the Angoff method task of judging item difficulty may be dependent on the experiential background represented by the participants, and on the PLD and item stimulus design used for the standard setting process. These generalizability concerns also follows from the fast and frugal heuristic perspective (Gigerenzer et al., 1999), according to which a heuristic strategy may not be generalizable to all contexts and may not yield accurate judgments in all circumstances.

Nevertheless, the findings from analysis of this practice round of judgment that followed training instruction and practice for both training suggested that even with deficiencies in knowledge and skills, the heuristic principles can still yield reasonable outcomes. The take away from evidence based on this practice round judgment data is that probability judgment is necessarily a reasoned process that involves recall of information and use of concepts with the concepts been neither dimensional nor metric. The evidence also extended the body of knowledge of standard setting by providing inductive support from the cognitive psychology heuristics and biases, fast and frugal heuristic, feature matching theory of similarity judgment knowledge bases (Gigerenzer et al., 1999; Kahneman & Tversky, 1974; Tversky, 1977).

B. Evidence for Transfer After Training Feedback

Although the Normative training group, based on all indicators, performed just about equally well as the Heuristic group in the practice round that followed training instruction and practice activities, the impact of the qualitatively different training was revealed in the feedback rounds. The Normative training groups recommendations correlated significantly much less with substantive domain DOK construct, correlated significantly higher with feedback norms, were less stable, were influenced more by group polarization phenomenon, and were more positively biased in relation to substantive evaluation considerations across feedback rounds. The findings of significantly declined relationship of item judgments with DOK, significantly higher correlation of judgments with feedback norms, higher although statistically insignificant positive bias of cut scores, and with-in participant probability judgment variability for the Normative training group across feedback rounds indicated higher sensitivity to and undesirable feedback influence. It was suggestive of that in the absence of appropriate training instructions that the prevailing feedback types that serve to present the participants with a possible representation of the final answer of judgment and without substantive considerations of the knowledge and skills requirements of the Angoff method tasks may have more undesirable influence.

Meanwhile, the observed impact of feedback regardless of training method in this study were as follows: (1) increasing correlation of participant's item difficulty judgments with item difficulty feedback norms, and decreasing standard deviation of correlations across rounds of judgment; (2) the participants rank ordering of items in terms of difficulty increasingly correlated higher with the empirical item difficulty ranks and their absolute item difficulty judgments with their study group mean estimates; (3) the cut score on the feedback test was higher even though in terms of substantive domain DOK consideration it was supposed to be more difficult; and, (4) the standard deviation of cut scores declined while on the average they were significantly positively biased in the feedback rounds in relation to the bootstrap PLD cut score cross validation criterion estimate, a finding consistent with a phenomenon called the group discussion induced polarization in the social psychology literature.

Based only on evidence of shrinking of standard deviations of correlations and increasing correlation measures across rounds in this study, it appeared that feedback had the intended effect and especially on the Normative training participants. However by relying also on rank order correlation, substantive domain theory, and distance metric based evaluation it was observed that the increasing correlation with feedback norms was at the expense of less substantively meaningful judgments. The evidence about the impact of feedback in this dissertation taken together suggested that participants' revision of their judgments due to feedback especially the Normative training group was not necessarily due to better learning. Rather their revisions were most probably motivated by conformity pressures to integrate feedback information and technical adjustments such as increasing or decreasing estimate to conform to feedback data. Therefore, the conclusion based on the evidence is that feedback did not have the intended effect and especially on the Normative training.

Notably the consistent finding of previous Angoff studies about the impact of feedback is that of increasing correlations of participant's judgments with feedback norms, and decreasing standard deviation of correlations across rounds of judgment (ACT, 1995a, 1995b; Clauser et al., 2002; Clauser et al., 2009a; Cross, Impara, Frary, & Jaeger, 1984; Hanick, 1999; Impara & Plake, 1998; Norcini, Shea, & Kanya, 1988; Reckase, 2000; Wyse, in press). The predominant conclusion by researchers so far based on these correlational and standard deviation indices have been that feedback does enhance judgment performance therefore the emphasis have remained on introducing different types of feedback into training and without instruction.

However, a more recent study by Wyse (in press) based on observation of large discrepancies of the item difficulty judgments at the presumed intended cut score of the participants despite the increased correlation suggested the alternative explanation of restriction of range for the phenomena of increasing correlation and called for re-examination of correlation as an index for evaluating Angoff method outcomes. The additional evidence about the impact of feedback provided by this dissertation based on substantive considerations lends support to

Wyse (2009) observation that correlation might not be the best index for evaluating Angoff method outcomes. The findings also indicated that range restriction was not the only factor driving the increasing correlations of participants judgments with criterion measures so that increasing monotonicity of judgments with feedback norms, and convergence of participants judgments influence mechanisms were also in play.

Other findings of analysis of the feedback rounds of judgment data which although were statistically insignificant but which may have practical relevance was that the Heuristic training participants cut scores maintained increasing trend across feedback rounds while the Normative training participants cut score declined in the second round of feedback that followed the construct map although was still higher. The possible explanation for the consistent increasing trend of cut score for the Heuristic training is because of discussion among the participants in all the feedback rounds of judgment that served to expose them to others judgment. This observed influence of group discussion feedback on final judgment outcomes does have inductive support from the social psychology literature where a robust phenomenon is the group discussion induced polarization effect (Fitzpatrick, 1989).

At the same time, the finding of decline in cut score in the second feedback round in the direction of the substantively determined PLD cut score for the Normative group generated the speculation that if the Normative training participants had received the construct map before the group discussion feedback, that there may not have been higher positive bias of their cut scores than those of the Heuristic training in the feedback rounds. Put differently, it may be that if the construct map feedback was given precedence and with appropriate instruction it would have mitigated undesirable influence of feedback.

8.3.4. Summary Interpretation of Findings

The desired influence of the training interventions in accordance with the goal of this dissertation of fostering veridicality of judgment outcomes should be enhancement of conceptual understanding, substantive reasonableness, and predictability of judgment outcomes. In accordance with this goal, it is desired that feedback have additional impact of bolstering conceptual understanding, substantive meaningfulness, and predictability of judgment outcomes. However, the evidence presented in this dissertation indicated that both training participants were obviously influenced by training interventions however, instruction and practice activities had more positive impact than feedback in relation to the afore stated goal.

Irrespective of training method instruction and practice activities appeared to engender more desirable influence on item difficulty judgment behavior than feedback. The finding of lowered substantive meaningful of cut score judgments of both training due to feedback was suggestive of undesirable influence of feedback. This finding of lowered substantive meaningfulness of the Heuristic training cut score judgment in the light of feedback despite the qualitatively better training instructions and practice activities was suggestive of need for re-consideration of the prevailing final outcome feedback types with instructions tailored to the knowledge and skills requirements of the Angoff tasks.

As Reckase (2001) suggested there is need for *a priori* specification of goals of standard setting training. As evidenced in this dissertation, the goal of training and evaluation approach were key consideration for validity of the conclusion drawn about training methods. Based on the findings of enhanced technical qualities of declining standard deviations and increasing correlations the conclusion, if the goal were merely fostering convergence of participants judgment to norms would have been that the Normative training was more effective than the

Heuristic training and that feedback was more effective than instruction and practice. However, in accordance with the goal of veridicality of this dissertation, although it is desirable that training interventions that includes feedback have as much influence as possible especially in yielding judgments that are comparable for participants, however it should not come at the expense of declined substantive meaningfulness of judgments. Hence, based on the observed decline of substantive meaningfulness of judgment outcomes in the feedback rounds, it was determined that feedback had undesired influence especially on the Normative training participants. Therefore, the reverse conclusions were reached that the Heuristic training was more effective than the Normative training and that training instruction and practice were more effective than feedback.

To conclude, the lack of clarity of goal of training and inappropriate choice of model for evaluation may have contributed to the gloomy picture about the capacity of the participants to render the Angoff method judgments, the cognitive complexity critique of the method, and the conclusion that the prevailing feedback types does help to remove inconsistencies in judgment. The projection is that reconsideration of the goal of laboratory standard setting and the Angoff method with better training instruction, practice, feedback, and comprehensive evaluation approaches as illustrated in this dissertation would facilitate accomplishing veridicality of standard setting outcomes.

8.4. Recommendations for Future Practice

The bases for recommendations made in this section are the findings of this dissertation empirical study. The inductive support from the cognitive psychology and the social psychology knowledge bases for the findings provides firm bases for making recommendations. Besides, the findings replicated previous Angoff method phenomena that spurred the cognitive complexity critique and feedback reliance, and unveiled additional findings.

The findings call into question the approach to research and the evidential basis for drawing the conclusion of the infeasibility of the Angoff method. Rather than suggest any problem with the Angoff method or with the participants, the finding indicated instead that the problem is with the approach to training and evaluating the performance of the participants. The evidence highlights the need for reconsideration of the Angoff method in the public school standard setting context with the Heuristic training method, better stimulus designs, training discussion designs, and evaluation criteria. Also, the Heuristic training for the Angoff method should be considered further with feedback tailored to the fundamental heuristic principles that would help participants to gain the type of conceptual understanding that is required for performing the Angoff method tasks.

There are obvious benefits that would accrue by reconsideration of the Angoff method. First, the Angoff method can yield more accurate cut scores than the Bookmark method which is currently the method of choice in the public school standard setting contexts. Second, the Angoff method can yield reasonable outcomes when participants are appropriately trained as evidenced in this dissertation so that it is a feasible method to use. Reconsideration of the Angoff method with the Heuristic training method in addition to the potential benefit of enhancing accuracy of judgment outcomes would also yield time and cost savings to standard setting enterprises.

For instance, as an under-resourced graduate student (no grant funding), it cost me roughly $4,000 to run two studies of 8 hours each on two separate days and yet the participant's judgments simulated what is obtained in practice with the extended training procedures. It is notable that the Heuristic training was implemented in this dissertation to replicate the traditional practice of including multiple feedback rounds. However, the findings of this study of reasonable judgments from participants following training instruction and practice activities suggested that feedback rounds of judgment may even be unnecessary which been the case would also lead to considerable reduction in implementation time for practical purposes. As a rough estimate the breakdown of cost for specific items in this dissertation implementation of the Heuristic training were as follows:

Food	$660.46
Facilitator	$450.00
Data Analyst	$150.00
Participants	$2200.00
Office Materials	$300
Printing	$200
Total	$3960.46

The room for this dissertation study of the Heuristic training was free of charge because it was conducted in a facility at the Michigan State University, the school of the researcher. The rest of the discussion in this section elaborates on recommendations for future consideration of the Angoff method with the Heuristic training method. The recommendations are organized in five topical areas namely: participants, discussion table allocation design, stimulus designs, training, and training evaluation in the subsections that follows.

8.4.1. Participants

The gold standard for future practice in the public school contexts would be recruiting teacher participants with adequate knowledge of the interaction of the target student population with test items and to augment for deficiencies in substantive domain knowledge through training instructions and practice activities. The expectation is that with adequate knowledge about students, far better results would be obtained with the Heuristic training[16]. Also, the finding of this study of considerable reasonable judgment performance in the practice round of judgment by the Normative training participants which comprised of a good number of math specialists, despite inadequate instruction and deficiencies in firsthand experience with the students highlighted that content expertise is another essential experiential characteristic of participants of standard setting. To the extent that the participants are recruited that have experience teaching students and that are content experts there may not be need for elaborate training on substantive domain content as required by the Heuristic training, so that instruction would focus on how to perform the Angoff method tasks.

However, in due consideration of the logistics that may be involved in recruiting teachers and content experts participants, if diverse stakeholders that include parents and community leaders are recruited as done by NAEP, the type of control used in this study of assigning participants to table groups to create balance in experience and background characteristics is recommended for best learning results. More on the strategy of distributing experiential and background characteristics in table groups for PLD discussion is presented in the immediately following discussion.

8.4.2. Discussion Table Allocation Designs

Practitioners would be better served to adopt the table allocation design of this dissertation. The suggestion is that if diverse stakeholders are recruited for the Heuristic training

[16] i.e., substantive domain knowledge, experience with the grade level for which cut score is being recommended, and empirical knowledge of the students' abilities and testing episodes involving the interaction of the students with test items

that participants should be assigned to table group for discussion that ensures balanced distribution of experience and background characteristics. To facilitate the table group allocation the standard setting organization should elicit participant background characteristics and use the information for assigning the participants to table groups for PLD discussions prior to the study. Some recommended background characteristics for table group allocations that were also considered in this dissertation includes the indicator of teaching experience, current position held, and gender. Other potential table group allocation experiential and background variables that could be considered for future practical standard setting include: number of years teaching, number of years in educational field, indicators of area of specialization, and experience teaching subject matter and at grade level of interest for standard setting. Allocating participants to table groups for PLD discussion also opens up a lot of prospects for designs to ensure adequate distribution of experience in table groups. For instance, future designs for table group allocation could be based on first matching participants by the elicited experiential variables and then randomly assigning participants to table groups within the matched groups. Knowledge of table groups of participants would also greatly facilitate evaluation of study outcomes.

8.4.3. Stimulus Design

There is utmost need for thorough pilot testing of items with the Heuristic training. It is important to use items that are adequate measures of the constructs and with minimal construct irrelevant features that could contribute to difficulty. Appropriate design and selection of test items with adequate alignment to PLD for Angoff studies would potentially enhance performance of the participants. Also, appropriate design and selection of test items and documentation of the content domain constructs they measure would facilitate evaluation of training outcomes. It is also essential that stimulus materials that include Angoff rating instructions be printed and handed out to the participants for ease of reference during the rating tasks. In contrast to simple oral review of the Angoff method instructions, the latter suggested practice would serve as a constant reminder to the participants about the judgment process.

8.4.4. Training

There are two recommended options for future consideration of the Heuristic training from the perspective of criterion referenced training tailored to the judgment heuristic principles. The first recommendation based on practical considerations of cost and efficiency in implementation would be that an organization can conduct the Heuristic Angoff training method with adequate training instructions and practice activities tailored to the judgment heuristic principles. The second recommendation if an organization is not too constrained by time and resources and based on consideration of accuracy is to conduct the Heuristic training augmented by at most two rounds of feedback. The recommended feedbacks are those that serve to reinforce student performance conception, categorical item content domain knowledge and skills constructs, and judgment strategies. This second option is also very feasible and will not accrue additional time and cost compared to that currently in use in practical contexts. In the paragraphs that follow elaborated recommendations for these two criterion-referenced to judgment heuristic principles training are discussed.

To begin with, the recommended ordering of the Heuristic training activities are as follows: (1) review of background information about study; (2) review of knowledge and skills measured by the test; (3) practice; (4) PLD discussion; (5) Instruction on the Angoff method; (6) rating items; and, (7) feedback. Recommendations for the more important of these training activities are discussed as follows.

(1) Review of knowledge and skills measured by the test - The content strand of the test items appeared to be the meaningful level of topical categorization of the items, therefore it is suggested that it be given priority in the review of test content. Also because the DOK appeared to be the most predictive construct of item difficulty judgments, it is important that review of the knowledge and skills measured by the test include DOK. Interactive DOK instruction with illustrative examples of test items designated at each level should also be considered in future practice, as it might even have substantial positive effect on the judgment performance of the participants.

(2) Practice - Categorization activities such as designating items to content domain constructs they measure are highly recommended as alternative to the practice activity of taking the test. The idea of categorizing items is synonymous to PLD review that is already in place as they are both aimed at concept learning. However, while categorizing items is meant to foster learning of item knowledge and skills constructs, PLD review is meant to foster learning of the student performance category.

Arguably, a pre-requisite for learning of student performance categories is been adept with the knowledge and skills constructs measured by the test items. The practice activity of designating items to content domain knowledge and skills constructs they measure that include the DOK would potentially enhance conceptualization of test items, foster knowledge of substantive domain item difficulty ordering, and learning of the target student population which in turn would enhance item difficulty judgment performance. Practice activity should also be considered of rank ordering of items in terms of difficulty. Tests of recall and recognition could be included during the training process to facilitate ascertaining the effectiveness of these practice activities. For example, the practice categorization and rank ordering of items can be based on a different test while the rating of items might be based on another with content matched for both tests to facilitate recall and recognition tests.

(3) PLD discussion- For PLD reviews the recommendation is that the elaboration process be maintained of asking participants to narrow down the descriptors to the knowledge and skills of the target group of students that barely make it in a performance category. In addition, the recommendation is that the participants could be instructed after describing the knowledge and skills of the students that barely make it in the performance categories, to elaborate further on the PLD by designating content strand and DOK of each of their descriptions and to rank order the descriptors in their order of difficulty for the target student population. This suggested additional PLD elaboration activity follows from the tenets of the feature matching theory (Tversky, 1977).

Also, because content strand and DOK appeared to be the psychologically meaningful constructs driving participant’s judgment implies that they might be better able to conceptualize the target students in terms of content domain constructs. The PLD elaboration activity of identifying the representative content strand and DOK of the knowledge and skills descriptions of the target students would facilitate feedback to the participants on their conceptions of the students. It would also facilitate use of the bootstrap PLD evaluation and the psychometric Rasch item response model approach of this dissertation to operationalization of the intended cut score of the participants[17].

[17] Refer to Reckase (2006) and Wyse (2009) psychometric theory approach to evaluation of standard setting outcomes for the notion of intended cut score

(4) Angoff method instruction – For Angoff studies using predominantly teacher participants the recommended instruction is the real with single or group formulations for conceptualizing the target students and items. Practical consideration of the real group of students and items formulation should instruct participants just as was done in this dissertation to think about what each item measures and to recall experienced items measuring same knowledge and skills constructs and to use them as the reference class for their judgment. An alternative prescription for practical consideration of the real and group formulation for conceptualizing test items when the test aligns well with the PLD may be to instruct the participants to match each test item to a PLD, recall and use discussed information as the basis for their judgment. If an organization is to consider the real single student formulation, the suggestion is that the equivalent second component task instructional formulation of asking participants to think about the proportion of times the student have been able to respond correctly to similar items as those on the test be prescribed.

For organizations that employ a diverse participant population and in general people who may not have firsthand experience with the target student population for which standards are been set, such as stake holders that include parents and community leaders as done with NAEP, the recommendation is that the hypothetical with single or group Angoff method instruction for conceptualizing the target students and items be considered. If the test items align well with the PLD, the hypothetical group instruction for conceptualizing the target students and test items should be considered with the instruction of matching each item on the test to a PLD, recalling, and using discussed information as the basis for item difficulty judgments.

Regardless of whether the real or hypothetical instruction is considered for the task of conceptualizing students and test items, practical implementations of the Angoff method should enforce instructions for the participants to weight highly the knowledge and skills constructs measured by the test items in their judgments. It is important to highlight that although considering unique features of items that impact difficulty is appropriate, however, they are considered unnecessary information for the participants to weight in their judgments, because, ideally they should be taken care of through item analysis procedures.

Thus, test items should be selected for Angoff studies that are adequate measure of the test constructs. It is also important to note that this dissertation studied the simplest case of Angoff standard setting in which cut score was sought for one performance category. The utility of training instructions on focused processing of construct information as a mechanism for reducing mental effort would become apparent in practical implementations where cut scores are sought for multiple performance categories. Moreover, standard setting for multiple performance categories is increasingly becoming the rule rather than the exception in the public school contexts especially with the *Race to the Top* alternative school accountability law which emphasis is on all performance categories as opposed to the Proficient performance category that is the focus of the NCLB law.

(5) Feedback – Because of the observed biasing influence of feedback on the cut score estimate of the Heuristic training in this dissertation, despite the qualitatively better training instructions and practice activities, group discussions on the final judgment outcomes of the participants are not recommended for further consideration. The finding

of lowered biasing influence of feedback on the Normative training participants' cut score judgment following the construct map feedback also indicated that the final outcome feedback types that serve to present multiple frames of reference are better than those that give participants a single frame of reference. Therefore, the recommendation is that practice consider of highest priority feedback discussions on intermediate outcomes and the elaborated construct map types of feedback augmented with instruction for the Heuristic training. Also, because of the finding of this dissertation of undesirable effect of mixed feedback types with the Heuristic training, the recommendation is that future practice considers sticking with one type of feedback during Angoff method training instead of the prevailing approach of providing different types across the rounds of judgment. For instance, an organization can choose to focus only on feedback discussions on intermediate outcomes and to keep reinforcing that throughout the training process.

The number one recommended feedback is group discussions led by the facilitator on intermediate outcomes that include student performance category, item categorical knowledge and skills constructs, and judgment strategies. The goal of feedback discussion with the Heuristic training is to reinforce student concepts, item concepts, and judgment strategy. Discussion of the knowledge and skills constructs measured by the test items can be based on graphical displays of the participant's designations of the items to content domain knowledge and skills constructs and content expert designations of the items just as was done for DOK in this dissertation. To help participants with better conception of the target student population that barely make in performance categories, substantive construct map could be generated for feedback based on operationalizing participants descriptions of the knowledge and skills of the target students and used to facilitate individualized instructions. To facilitate this substantive construct map feedback, preparatory work should be done of selecting example items for training that measures the PLD. Items should then be selected for creating the construct map that measures the knowledge and skills of the target student population as described by the participants during the PLD review. The substantive construct map should be generated for the test used for the Angoff ratings at ability levels that correspond to the Rasch model difficulties of the items that measures the knowledge and skills of the target student population as described by the participants.

The Rasch model is recommended for generating the substantive construct map because it is the closest item response model to the judgment heuristic principles. In addition to specifying test items Rasch model difficulties and cut score of test at the ability levels that correspond to the difficulties of the selected items that measure the knowledge and skills of the target students as described by the participants, the substantive construct map should also delineate the DOK and content strand of the items. The substantive construct map feedback would serve to give the participants a sense of the possible response profiles of the target student population that they described during the PLD review.

Fostering consensus among participants on the knowledge and skill constructs measured by test items and on the conceptualization of the student performance, coupled with complete explication of the Angoff method instruction would also help to achieve the goal of variance reduction of item difficulty and cut score judgments. This goal of variance reduction is espoused by the current practice, however achieved instead through group discussion of the final judgment outcomes namely, the item difficulty and cut score judgments which process is at the risk of biasing the revised estimates.

The second recommended feedbacks are the prevailing final outcome construct maps that are created external to the training environment and without reference to the *a priori* conceptions of the participants, and the unconditional empirical item response data with instructions on how to integrate them into judgments. For the final outcome construct maps, the participants can be instructed to use the data as reference points and to adjust recommendations of greater uncertainties if they want to and to align with a particular cut score level while maintaining substantive domain knowledge and skills categorical consistency. The unconditional empirical item difficulties based on student responses can also be considered for feedback with instruction tailored to the Bayes rule. The instruction could also be considered with the unconditional empirical item difficulties, to revise ratings of greater uncertainties while retaining item content domain knowledge and skills categorical consistency.

It is essential that feedback designs continuously assess acquisition of the knowledge and skills requirements of the Angoff item difficulty judgment task and in accordance with the judgment Heuristics principles to facilitate evidence that any observed changes in participants' judgments are due to desired effect of better cognitive learning. Therefore, feedback should be intertwined with formative assessments of the pre-requisite knowledge and skills of recall, recognition, categorization, and relative item difficulty judgments. Continuous criterion-referenced assessment of performance due to feedback would also provide information that would enable building validity evidence about the effectiveness of feedback.

8.4.5. Training Evaluation

The comprehensive evaluation approach of this dissertation based on the Kirkpatrick's training measures framework, the substantive domain bootstrap resampling method, and the multidimensionality reduction analytic framework are highly recommended for future practice. Use of multiple evaluation approaches would offer more insights about the effectiveness of training interventions and also enhance the validity of conclusions.

Evaluation measures should encompass the pre-requisite knowledge and skills requirements of recall, recognition, categorization, and difficulty rank ordering of items. It is also essential to measure non-cognitive attributes of participants that could potentially confound with training effectiveness such as motivation, engagement, and emotion. Self-report measures should also be considered further that includes questions on confidence, understanding, satisfaction, factors considered in judgment, and strategies that the participants applied in their judgments.

The bootstrap resampling method should be considered for operationalizing participants' PLD discussion descriptions of the knowledge and skills of the students that barely make it in a performance category for evaluation purposes. Items should be selected that measure the PLD review knowledge and skills descriptions of the participants, with the bootstrap resampling method applied to the Rasch model difficulties of the items to estimate a representative cut score on the IRT scale and test items difficulties. The Rasch model is recommended because it is the closest item response model to the judgment heuristics principles. Cross validation check of participant item difficulty and cut score judgments with the bootstrap estimates would be a substantive check of the consistency of the participants' judgments with their intended estimates[18]. The PLD can also be operationalized with the bootstrap resampling method as was demonstrated in this dissertation for cross validation check of participant's judgments however with many experts' item selections that measure the knowledge and skills of the PLD.

The multidimensionality reduction analytic framework is strongly recommended for evaluation of Angoff method judgments. For Angoff studies based on small sample sizes of tests and participants as was the case for this dissertation, it is recommended that principal coordinates analysis be supported by parallel analysis to determine the characteristics of results related to the sample size and the validity of the solutions.

8.5. Study Limitations

A number of knowledge and skills assumptions about participants were relaxed and that are testable via evaluating the impact of training interventions that served to reinforce them. One, the participants do not have prior knowledge about the DOK so that training instruction that reinforces this knowledge would enhance judgment performance. Two, practice categorizing test items constitutes more meaningful processing of the knowledge and skills measured by the test than taking the test and would improve categorization, recall, and judgment performance. Three, participants are better able to conceptualize real students than hypothetical students so that training instruction that reinforces it would enhance judgment performance. Four, participants do not know how to integrate knowledge and skills constructs measured by the test items into judgment of item difficulties with the Angoff method, so that training instruction that reinforces judgment strategy would improve judgment performance. Five, deep understanding of knowledge and skill constructs measured by the test, knowledge of and conceptualization of real

[18] Confer Reckase, 2006; Wyse, 2009 for a psychometric theory based evaluation assumption.

students, and knowledge of judgment strategies would enhance item difficulty judgment performance. Six, feedback discussion of the constructs measured by the test and reinforcement of judgment strategies would enhance substantive meaningful of judgment process and outcomes net the impact of the interventions of instruction and practice. However, because this dissertation is a small scale study it was not possible to test each of these knowledge and skills assumptions and the effectiveness of each of the training procedural interventions directed at reinforcing them.

Therefore, the limitations of this dissertation pertain to the short test lengths, small sample sizes of participants, less than ideal statistical controls, and the simplicity of the study design used for the test of research hypotheses. Because of the small sample sizes involved, there was insufficient power for statistical tests of significance so that there may be characteristics of results of statistical analysis procedures related to the sample size. In addition, constraints in embarking on ideal statistical controls and the simplicity of the study design did not allow for complete accounting of all potential confounding variables, accounting of training intervention order effects, and for singling out the effects of individual training interventions of instruction, practice, and feedback.

These limitations underscores the need for large scale studies with better designs such as pre-post tests and counterbalancing designs to further test the knowledge and skills assumptions and the effectiveness of the training interventions aimed at reinforcing them. Also, small scale studies employing the two-way mixed effects ANOVA design of this dissertation can test further individual knowledge and skills assumptions and the effectiveness of the specific training interventions of instruction, practice, and feedback aimed at reinforcing them through varying the specific intervention of interest for the involved groups while keeping all else same.

Nonetheless, regardless of the sample sizes, statistical control, and experimental design limitations of this dissertation, the findings have inductive support of the judgment heuristic and social psychology knowledge bases. The findings also replicated and clarified previous Angoff standard setting training intervention research phenomena. In the ensuing discussions in Section 8.6., suggested visions and directions for future Heuristic training research are presented. For each of the suggested future lines of the Heuristic training inquiry, it is recommended that evaluation of training effectiveness consider comprehensive training evaluation framework as demonstrated in this dissertation.

8.6. Directions for Future Studies

There are two philosophical views that could be adopted for studying further the Angoff method Heuristic training instructions namely: (1) the operationalist measurement, the subjective probability, and the value judgment view of standard setting; and, (2) the realist measurement, the objective probability, and the parameter estimation view of standard setting. The philosophical view adopted for this dissertation was the realist measurement, the objective probability, and the parameter estimation view of standard setting. In accordance with the realist measurement, the objective probability, and the parameter estimation view of standard setting, teachers are the ideal participants of Angoff studies. Hence the Angoff method instructional formulation tested by this dissertation for the ideal teacher population was conceptualization of real group of students and items. This instructional formulation to conceptualize real group of students and items was compared to instruction to conceptualize hypothetical group of students and items with the latter instruction representing the prevailing operationalist measurement, the subjective probability, and the value judgment view of standard setting.

There is inconclusive evidence based on this dissertation study findings about the impact of the two instructional variants and for the ideal teacher population. Future studies with better design controls, idealized teacher participants with firsthand knowledge of the interaction of students with test items can test further the assumption that conceptualizing a group of real students is simpler and can yield more accurate judgment than that of conceptualizing a hypothetical group by varying only this Angoff method instruction part for the two groups while keeping every other aspects same. There is also need for further studies comparing Angoff method instructions conceived of based on these two philosophical views with diverse participant pool.

Comparative Angoff method philosophical view instructional formulation studies can test further the cognitive simplification hypothesis through variations of instructions on conceptualization of a single versus group and/or a hypothetical versus real students and with corresponding formulation for conceptualizing test items for the task of estimating item difficulties. The studies can also test Angoff instructions that specify the group size of real or hypothetical students and items to conceptualize. Different specifications of the number of content domain knowledge and skills features of items to consider in conceptualizing test items can also be tested for effectiveness by research studies comparing the hypothetical versus the real student and item Angoff method instructions. Such studies can also consider judgment reaction time measures, in addition to judgment accuracy measures to adequately test the cognitive complexity reduction hypothesis.

There are four broad visions for future research studies choosing to focus on understanding of Angoff method Heuristic training conceived of based on either one of the aforementioned two philosophical views namely: (1) studies of effectiveness and generalizability of Angoff method Heuristic instructions; (2) studies of effectiveness and generalizability of practice activities; (3) studies of effectiveness and generalizability of feedback procedures; (4) Evaluation focused research based on criterion referencing to the judgment heuristic principles. Each of these four visions for future research is elaborated upon in the subsections that immediately follow.

8.6.1. Studies of Effectiveness and Generalizability of Heuristic Instructions

Regardless of the philosophical view adopted for Angoff method instruction, there is need for studies to compare different types of DOK and content domain knowledge and skills constructs instructions such as interactive versus non-interactive to investigate which way of communicating the information works best. There is also need for studies investigating different types of PLD elaboration instructions to assist participants with the conception of real or hypothetical students that fit the PLD in order to minimize influence of *a priori* biases. For instance, studies can test the effectiveness of describing the students that barely make it in performance categories augmented with identifying the representative DOK and content strand measured by the PLDs as opposed to just generating the descriptions. PLD instruction can also be tested that uses the construct map that delineates the content strand, DOK, and example items measuring the participants knowledge and skills descriptions of the target students to facilitate discussions. Future studies can test either one of the instructional formulation of the Angoff tasks of real versus hypothetical with different PLDs, test stimulus design contexts, experiential background of the participants, PLD discussion, and experimental design arrangements.

Studies adopting the philosophical perspective of the realist measurement, the objective probability, and the parameter estimation view of standard setting should consider making a priority, investigating the instructional formulation to conceptualize a group of real students.

Such studies, prescribing instruction to conceptualize a group of real students could consider the second component task instruction of thinking about the proportion of the students that would respond correctly to the test items with teacher participants.

It is important to highlight that the Angoff method task was conceived of for this dissertation based on the realist measurement, the objective probability, and the parameter estimation view of standard setting as a two-step process namely: (1) conceptualizing the target students; and (2) estimating item difficulties for the students. However, in hindsight, the Angoff method task is more appropriately conceived of based on the realist measurement, the objective probability, and the parameter estimation view of standard setting as comprising of at least a three-step process that entails: (1) conceptualizing the target students; (2) conceptualizing items measuring knowledge and skills similar to those on the test; and, (3) estimating difficulties of items on the test. In accordance with this re-conceptualization of the number of steps involved in the Angoff tasks implies that the Angoff method instruction in this dissertation on how to estimate item difficulties once the target student group is conceptualized was an improvement over the prevailing instruction, only in terms of specification of how to conceptualize items. Hence, the Angoff method instruction on how to estimate item difficulties was still unexplicated as to whether recall of previously experienced similar items should be explicit, how many items to recall, and on how to combine estimates in the circumstance that more than one item is recalled.

On this note, it is important to re-affirm that the goal of adopting the realist measurement, the objective probability, and the parameter estimation view of standard setting philosophical view for the ideal teacher participants and for prescribing conceptualization of real group of students and items was to make the task of estimating item difficulties more objective. Future studies should explore further explication of the task of judging item difficulties for the real and group formulation and the effectiveness of different instructional specificities with teachers. In addition to explicating the content domain knowledge and skills categories to consider in judgment deliberations, such future studies could also consider specifying how many real students and real test items to recall and if more than one experienced test item is recalled, how to combine the relative frequency estimates.

Studies directed at understanding the real group versus hypothetical group of students and items Angoff method instructional formulations would illuminate the best subset of Angoff method heuristic strategy instructions that can effectively reduce cognitive complexity of the tasks while maintaining the goal of veridicality when the participants are knowledgeable about the interaction of students with test items and that can be carried out in a reasonable timeframe. To summarize, this future line of Heuristic training inquiry should investigate further, the possible effects on recovery of the heuristic model assumptions when using participants with different levels of experience with the student population, different discussion table designs and matching variables, tests comprised of items with different degrees of match with PLDs, with the Angoff method instructions. The goal of this line of inquiry regardless of philosophical view adopted should be to understand the circumstances in which the instructions are most effective.

8.6.2. Studies of Effectiveness and Generalizability of Practice Activities

Regardless of philosophical view adopted for Angoff method instruction, one possible training study in this category that immediately comes to mind is comparison of the practice activities of test taking versus designating items to knowledge and skills constructs they measure to investigate that one improves recognition, recall, and judgment performance over the other. There are two possible hypotheses that could be tested for the practice activity of designating items to knowledge and skills constructs that they measure by such future studies (1) that it improves categorization and recall performance and, (2) that it improves performance in the Angoff judgment tasks.

To test both hypotheses while separating out the effect of practice from Angoff method task instruction in a future study, the recommended ordering of training activities is to review the knowledge and skills measured by the test items then have one group take the test and the other group to do the practice activity of categorizing items. Both groups could then be tested for recall and recognition performance. Subsequently, the groups can elaborate on the PLD, followed by unexplicated Angoff method task instructions about how to judge item difficulties then, the participants would judge item difficulties. To test also the effect of the explicated Angoff method task instruction in the same study, subsequently one group could then be given explicated instruction and the other unexplicated instruction about how to perform the task of estimating item difficulties, then the participants would then repeat judging item difficulties.

There are a lot of creative options on how to test separate hypothesis about the Angoff method task instruction versus training practice activities. The most important thing to note by such future studies is to ensure that the two interventions of taking the test versus categorization should take place in the same positional order in both training. The recommendation is that practice take place right after the discussion of the knowledge and skills constructs measured by the test and before the PLD review. Future studies can also compare a training in which participants perform both practice activities of categorizing and taking the test to a study in which the participants perform only one of the activities to test if there is differential effect on the judgment outcomes.

8.6.3. Studies of Effectiveness and Generalizability of Feedback Procedures

Regardless of the philosophical view and the formulation of the Angoff method task instruction, there is need for studies with better experimental designs to test for the net effectiveness of the intermediate outcome types of feedback based on continual reinforcement of the knowledge and skills constructs measured by the test, the performance categories, and judgment strategies through group discussions. Such studies can explore the model of feedback used with the National Assessment of Educational Progress (see report by Raymond & Reid, 2001). The NAEP model is based on iterative rounds of feedback discussions. The proposal is that in addition, the studies be based on continual reinforcement of content domain knowledge and skills constructs, items and person performance categories, and judgment strategies. These studies based on iterative reinforcement of item and performance categories, and heuristic strategies should also incorporate formative assessment of the knowledge and skills of participants as a means of validating the effectiveness of the feedback mechanisms.

The substantive construct map feedback can be compared to the intermediate outcome feedback discussions. The substantive construct map can be generated for the test items at ability levels corresponding to the Rasch model difficulties of items that measure participants' descriptions of the knowledge and skills of the target students. The accompanying instruction can emphasize that participants maintain item content domain categorical construct consistency in adjustments to item judgments of greater uncertainties. Also, studies can compare the prevailing construct map feedback type that are generated without reference to the *a priori* knowledge and skills descriptions of the participants in the training environment and with instructions to maintain content domain construct consistency in making revisions to items versus the intermediate outcome feedback discussions. Future studies can compare the unconditional empirical item *p*-value with instruction to participants to maintain item content domain categorical consistency in adjustments to item judgments of greater uncertainties or with instructions tailored to the Bayes rule versus the intermediate outcome feedback group discussions.

Large scale studies considering all three types of extended feedback training namely, the intermediate outcome group discussion feedback, the substantive and/or final outcome construct map feedback with instruction, and the unconditional empirical item p-value with instruction could be based on counterbalancing designs to adequately control for order effects. Such studies can also incorporate formative assessments of content domain knowledge and strategies to augment validity evidence that these feedback operations lead to better conceptual understanding and improvement in substantive meaningfulness of judgment process and outcomes.

8.6.4. Evaluation Focused Research

There is need for evaluation focused research to re-analyze existing Angoff method data using the multidimensional scaling and cluster analytic techniques. If the information still exists, the bootstrap resampling approach of this dissertation can also be used to operationalize the knowledge and skills descriptors of the participants during the standard setting as a substantive test of the assumption of the psychometric theory of the concept of intended cut score (Reckase, 2006 & Wyse, 2009). The bootstrap resampling approach to operationalizing the PLD can also be used by these studies with multiple selections of items measuring the knowledge and skills of the PLD by different content experts. Such research can also explore other approaches to operationalizing the knowledge and skills of the PLD for cross validating Angoff method outcomes. Other alternative approaches to cross validating the Angoff method study outcomes can also be mechanistic and field approaches solutions and in relation to criterion referencing to item knowledge and skills constructs.

8.7. Conclusion

The findings of this dissertation were that for both the Heuristic and the Normative training methods, instruction and practice activities yielded substantive meaningfulness and reasonableness of judgment outcomes. In contrast, feedback enhanced the technical qualities of correlation and standard deviation especially for the Normative training but was to the detriment of substantive meaningfulness and reasonableness of judgment outcomes. The conclusions drawn based on the findings were that the Heuristic training was more effective than the Normative training and that instruction and practice were more effective than feedback.

The findings about the impact of training instruction and practice addressed McGinty’s (2005) call to understand the factors and cognitive processes of standard setting tasks and built on standard setting knowledge bases of the cognitive processes underlying standard setting tasks and effectiveness of training interventions. Moreover, the evidence provided by this dissertation about the impact of training instruction and practice indicated that the critique of cognitive complexity of the Angoff method would be discredited with appropriate training and also underscored the need for better design of feedback.

Because the study findings have inductive support from the cognitive psychology judgment heuristic and the social psychology literatures, it was firm bases for making recommendations for future standard setting practice. The recommendations were that public school standard setting practitioners should re-consider the Angoff method with the Heuristic training, with better test stimulus designs, table group allocation designs, and appropriate feedback that reference the knowledge and skills requirements of the tasks.

By replicating and answering important questions about existing controversies in standard setting literature about the Angoff method through adequate reliance on theory in the design of the training program, this dissertation extends the broad literature on standard setting while establishing groundwork for further investigations of the Heuristic training with better statistical designs. The evidence provided by this dissertation reframes the debate about the utility of the Angoff method, about the approach to training, and evaluating training programs. It is hoped that this dissertation would stimulate research on the design of other standard setting methods training and based on substantive theory considerations. The broader impact of research efforts targeted at the design of standard setting methods training are discussed as follows.

The design of training based on criterion referencing to the knowledge and skills requirement of standard setting methods tasks has the potential of improving the performance of participants in the tasks and increasing the accuracy of standard setting outcomes. Increase in the accuracy of standard setting outcomes would ultimately reduce errors in classifying students to performance categories. Moreover, emphasis on training instruction and practice activities would potentially increase knowledge, improve skills, and change attitudes of the target classroom teacher participants. These changes when transferred to their classroom contexts would greatly enhance their classroom practices and ultimately facilitate teaching professional development. In addition, increase in the accuracy of standard setting outcomes would ultimately help in identifying schools needing intervention with the public school accountability programs. Consequently, proper identification of schools needing intervention would promote the goal of the public school accountability system of school organizational improvement, increase teaching practices, and student learning. These changes would in turn ensure that the best brains practice the professions and impact the society at large.

APPENDICES

Appendix A:

Definitions of Key Concepts

Angoff Method
A standard setting method with task requiring participants to judge conditional probabilities of correct response to dichotomously scored test items for a student population that barely makes it in a performance category

Availability
A heuristic process applied by humans in probability judgment that entails recall of what easily comes to mind

Bookmark Method
A standard setting method in which items are ordered in terms of item response difficulties with task, requiring participants to place a bookmark on the first item in the booklet that the student population that barely makes it in a performance category cannot answer correctly with the booklet ordering probability criterion

Categorization
A cognitive process that entails identifying the conceptual groupings of objects

Cognitive Task Analysis
Methods for understanding the knowledge and skills underlying complex task performance

Cutscore
A point on the test score scale used for classifying students into performance groups based on their demonstrated knowledge and skills competencies

Heuristic
A simplified strategy applied by humans in the judgment of probability that ignores part of the information with the goal of making judgment more quickly, frugally, and accurately

Judgment
A slow and deliberate thinking process of integrating information and for estimating the probability of events

Memory
A store house of human experiences and knowledge

Modified Angoff method
An operational variant of the Angoff method

Performance Standards
Qualitative descriptions of intended knowledge and skills a test taker must have to be classified at a particular performance level

Probability
The relative frequency of occurrence of an event

Rationality A capacity to reason correctly

Recall A measure of memory of experienced events

Representativeness A heuristic process applied by humans in probability judgment that entails judgment of similarity between an object and a category

Standard Setting An organized system for collecting the judgments of qualified individuals about performance standards and for translating knowledge and skills descriptions to cut scores on the test scale

Appendix B:

Scripts

The scripts are presented in this section and in the order of scheduled training activities.

Script B-1: Heuristic Training Script for Review of Background of Tests

[Hand out concept sheet, content strand, Grade Level Expectation and DOK review sheets then turn to the background: Subset MEAP Tests slide and say the following:]

I just gave you a concept sheet, content strand, third Grade Level Content Expectations (GLCE) and Webb's Depth of Knowledge (DOK) review sheets.
The concept sheet is meant to give overview definitions of key concepts we will be using in this workshop.

The practice and real tests that you will be working with are fourth grade mathematics. There are 15 items in both tests comprising of carefully selected subset 2005 released MEAP multiple choice items.

The term "Items" as we will be using in this workshop simply refers to test questions. Keep in mind that prior to their official use in MEAP tests, that items go through rigorous process of pilot testing so that the quality of the items in the practice and real tests is not in doubt.

The content strand, GLCE and DOK review sheets delineate the knowledge and skills precisely measured by both the practice and real tests that you will be recommending cut scores for in this workshop.

We will briefly review the information contained in each of the review sheets. However, it would also help if you could take time during breaks to digest the information contained in them as you will find them as useful references for your tasks.

Four content strands are represented in both practice and real test and are: Number and Operations, Measurement, Geometry and Data and Probability. The topics covered in each of the strands are specified in the content strand review sheet.
The practice and real tests are fourth grade mathematics, while the Grade Level Content Expectations measured is third grade. The reason for the latter as you may already know is that the MEAP test is administered in the Fall of each year.

For practical reasons that include making the task feasible, the practice and real tests measure only 13 of the third grade content expectations assessed by the complete MEAP test. The third grade content expectations measured by the items in both the practice and real test are specified in the Grade Level Content Expectations hand out.
In addition to GLCE's and content strands, the practice and real tests also measure reasoning skills that are based on the Webb's Depth of Knowledge levels. The DOK of an item is related to but is distinct from its difficulty. Please refer to the concept sheet for definitions.

According to Webb, there are four levels of thinking skills measured by items, these are: recall, skills and concepts, strategic and extended thinking. The DOK review sheet summarizes most of what you need to know about DOK levels.

The first column of the table gives the name of the DOK level, the second its definition as given by Webb and the third an example item at that level.
[Briefly highlight the distinguishing features of the DOK levels and ask them to refer to the hand out for details. Then say the following:]

Read and digest the DOK review sheet as you will find the information useful in your tasks. Any questions about the content of the tests you'll be working with or about where they came from or about any of the things we covered so far?
[If there are address them before continuing with the PLD discussion.]

Script B-2: Normative Training Script for Review of Background of Tests

[Hand out the content strand and Grade Level Expectation sheets. Turn Background Subset MEAP Tests slide and say the following:]

I just gave you content strands and third Grade Level Content Expectations (GLCE) sheets.

The practice and real tests that you will be working with are fourth grade mathematics. There are 15 multiple choice items of varying difficulty in both tests. The items comprise of carefully selected subset 2005 released MEAP items.

The term "Items" as we will be using in this workshop simply refers to test questions. Keep in mind that prior to their official use in MEAP tests, that items go through rigorous process of pilot testing so that the quality of the items in the practice and real tests is not in doubt.

The content strand and GLCE sheets delineate the knowledge and skills precisely measured by both the practice and real tests that you will be recommending cut scores for in this workshop.
Four content strands are represented in both practice and real test and are : Number and Operations, Measurement, Geometry and Data and Probability. The topics covered in each of the strands are specified in the content strand handout.

The practice and real tests are fourth grade mathematics, while the Grade Level Content Expectations measured is third grade. The reason for the latter as you may already know is that the MEAP test is administered in the Fall of each year.

For practical reasons that includes making the task feasible, the practice and real tests measure only 13 of the third grade content expectations assessed by the complete MEAP test. The third grade content expectations measured by the items in both the practice and real test are specified in the Grade Level Content Expectations hand out.

Any questions about the content of the tests you'll be working with or about where they came from or about any of the things we covered so far?

Script B-3: Heuristic Training Script for Modified Angoff Instruction

[Hand out the modified Angoff Procedure item rating instructions and say the following:]

You will be using the modified Angoff approach to recommend cut scores for proficient performance on the MEAP tests. The Angoff procedure requires rating test items. The ratings are in terms of percentages from 0 to 100 of the barely proficient students that would respond correctly to each of the items on the test. Your task is to generate this percentage for each item on the test while ours is to convert your recommendations into cut score and to give you feedback on them. The complete instructions to guide you with using this procedure to rate the items are contained in the modified Angoff rating instructions hand-out I just gave you, are:

A. Think about the barely proficient students

For each item on the test:

B. Think about what it measures (Content strand, GLCE, and DOK level)

C. Think about items that measure these same knowledge and skills

D. Recall or imagine the proportion of students who are barely proficient that would respond correctly to items in this category.

E. Mark the percentage from 0 to 100
Any question?
The overarching expectation is that you all rely on information provided in this training in giving your best judgment. However, for those of you with experience teaching fourth grade students, these instructions are really aimed at activating your knowledge of interaction of actual students in your classrooms, who match the descriptions of barely proficient with similar test items. Keep in mind that similarity of test items is defined for our purpose in terms of matching content strand, GLCE and DOK level. By all means, refer to the paper versions of these instructions and all materials provided in this training in making your judgments. Use your best judgment to make these decisions, but do not agonize over them.

Script B-4: Normative Training Script for Modified Angoff Instruction

[Hand out the modified Angoff Instructions and say the following:]

You will be using the modified Angoff approach to determine cut scores for proficient performance on the MEAP tests. The Angoff procedure requires rating test items. The ratings are in terms of percentages from 0 to 100 of the barely proficient students that would respond correctly to items. Your task is to generate this percentage for each item on the test while ours is to convert your recommendations into cut score and to give you feedback on them. The complete instructions to guide you with using this procedure to rate items are contained in the modified Angoff rating instructions hand out that I gave and are:

A. Imagine or think about a classroom made up of 100 barely proficient students

For each item on the test:

B. Based on description of barely proficient students, what proportion of the students in the above classroom would answer the item correctly?

C. Mark the percentage from 0 to 100

Refer to the paper versions of these instructions and all materials provided in this training in making your judgments. Give your best informed judgments but do not agonize over them.
Any questions?

Appendix C:

Instruments

The instruments used for the studies are presented in this appendix and in the scheduled order of the training activities.

Instrument C-1: Panelist Information Sheet

Panelist Identification Number: ___________________________________
This questionnaire should take no more than 10 minutes to complete. Your answers will be kept confidential. Please endeavor to **turn in this questionnaire on the day of the workshop.**
There are two parts to this questionnaire. The first part asks questions about your background while the second elicits your reasons to participate in the study.

PART 1

Please answer the following questions about your background.

How many years have you worked in the field of education? _________
What is your current position? __________________
How many years in current position? __________________
What is your area of specialization? __________________
What is your gender? □ F □ M
How best describes your ethnicity □ African American/Black
□ Asian/Asian
□ American/Pacific Islander
□ Caucasian/White
□ Hispanic/Latino
□ Other:_________________
If not currently a teacher, have you ever taught?
_____Yes
_____No
If yes, for how many years? _______________________
Which levels/grades have you taught or do you teach? _______________ Levels/Grades
Which subject areas ? ________________ Levels/Grades
How best describes your school district(s) □ Urban
□ Suburban
□ Rural

Instrument C-1: Panelist Information Sheet (cont'd)

PART 2

We are interested in learning about your reasons for participating in the workshop. Please rate each of the following statements. There are no “right “ or “wrong” answers. Choose the rating that best reflects your true reason.

Rate how true the following are for you by marking the cell that most describes your reasons for participating in this project. I am participating in this workshop ...

	Not At all True	**Somewhat True**	**True**	**Very True**
Because it's what was asked to do.				
Because it's important to me to acquire professional skills and to improve my career prospects.				
Because I will get remunerated.				
Because I love learning new things and improving my skills.				
Because I want to help the researcher.				
Because I will feel bad if I do not avail myself of this opportunity.				
Because I think I will acquire materials useful for professional practice.				
Because I am always seeking opportunities to acquire new knowledge.				

Instrument C-2: Heuristic Training Practice Questions

Instructions

Please make sure to write your unique panelist identification on the rating sheet. Please note that your identification number is used for analysis purposes only and that your responses to questions will be held in strict confidence. The questions are about the items in the practice test booklet. Please address the questions as completely as possible as this will help in providing feedback to further assist you in your task.

Please address the following four questions for each of the 15 items in the practice test booklet. The codes you are to use are shown after each question. Your responses to the questions for each item should appear in a single row of the rating sheet, corresponding to the position of the item in the test booklet.

Questions

1. What is the content strand and GLCE of item?

    The code for content strands in parenthesis are as follows: Number and Operations (N); Geometry (G); Measurement (M) and; Data and Probability (D).
    The code for a GLCE should be its number in the GLCE hand out.
    The content strand code of an item should precede its GLCE code. For example, if a measurement item and GLCE 1, your answer to this question should be M1.

2. What is the most likely DOK level of the item?

    The code for DOK levels in parenthesis are as follows: Recall (RE); Skills and concepts (SC); Strategic thinking (ST) and; Extended thinking (ET)

3. What is your best estimate of the proportion of students in the barely "proficient" category that would respond correctly to the item?

    This 3rd question pertains to the modified Angoff standard setting procedure. Mark the percentage from 0 to 100

4. What is the difficulty rank of the item with respect to other items in the booklet.

    This 4th question asks for rank order of the items in terms of difficulty. Please note that you should address it when you've gone through all items in the booklet. Rank items in terms of difficulty using a 1 for the **least difficult** and 15 for the **most difficult** item.

Instrument C-3: Normative Training Practice Questions

Panelist Identification Number___________________________

Instructions

Please make sure to write your unique panelist identification number on the rating sheet. Please note that your identification number is used for analysis purposes only and that your responses to the questions will be held in strict confidence. The questions are about the items in the Practice Test booklet. Please address the questions as completely as possible as this will help in providing feedback to further assist you in your task.

Please address the following two questions for each of the 15 items in the Practice Test booklet. The codes to use are shown after each question. Your responses to the questions for each item should appear in a single row of the rating sheet, corresponding to the position of the item in the test booklet.

Questions

1. What is your best estimate of the proportion of students, who match the description of barely "proficient" that would respond correctly to the item?

This 1st question pertains to the modified Angoff standard setting procedure. Mark the percentage from 0 to 100

2. What is the difficulty rank of the item with respect to other items in the booklet?

This 2nd question asks for rank order of the items in terms of difficulty. Please note that you should address it when you've gone through all items in the booklet. Rank items in terms of difficulty using 1 for the **least difficult** and 15 for the **most difficult** item.

Instrument C-4: Real Test Rating Questions

Please note that these same real test rating questions were addressed by the Heuristic and Normative Training. Also note that only questions 3 and 4 were addressed by both groups for the second round of real test judgment)

Real Rating Questions

Instructions
Please write your unique identification number on the rating sheet.
Please address the following four questions for each of the 15 items in the real test booklet. The codes you are to use are shown after each question. Your responses to the questions for each item should appear in a single row of the rating sheet, corresponding to the position of the item in the test booklet.

Questions
1. Does the item look familiar?
1 = yes 0 = no

2. Have you seen it before?
1 = yes 0 = no

3. What is your best estimate of the proportion of students in the barely "proficient" category that would respond correctly to the item?
This 3rd question pertains to the modified Angoff standard setting procedure. Mark the percentage from 0 to 100.

4. What is the difficulty rank of the item with respect to other items in the booklet.
This 4th question asks for rank order of the items in terms of difficulty. Please note that you should address it when you've gone through all items in the booklet. Rank items in terms of difficulty using a 1 for the **least difficult** and 15 for the **most difficult** item.

Instrument C-5: Heuristic Training Discussion Guidelines

Instructions

Please make sure that your voice is heard during this discussion

Do not let your partners control the discussion

Always keep in mind that there is no right or wrong answer.

Endeavor to give the rationale for your recommendation

Focus on the item judgments for which there are greater grey areas and/or disagreements

You are encouraged to take notes during the discussion and to begin to formulate necessary changes to your round one ratings if need be

Specific Issues to Address

Please pay attention to the following issues:

Content strands and Third Grade Expectations of items

DOK level of items

Difficulty ordering of items

Recommendation of proportion of barely proficient that would respond correctly to items
Strategies used for recommending proportion of barely proficient that would respond correctly to items (with focus on the modified Angoff strategy)

Instrument C-6: Normative Training Construct Map Feedback

Item Number	1	2	3	4	5	6…	15	
[redacted]								
Ability Level	**Item P-Values**							**Cut Score**
[redacted]	0.45	0.11	0.31	0.52	0.43	0.19	0.08	2.96
[redacted]	0.52	0.14	0.38	0.60	0.50	0.24	0.10	3.60
[redacted]	0.65	0.21	0.50	0.71	0.62	0.34	0.16	4.83
[redacted]	0.72	0.27	0.59	0.78	0.70	0.43	0.21	5.81
[redacted]	0.73	0.28	0.60	0.79	0.72	0.44	0.22	5.98
[redacted]	0.75	0.30	0.62	0.80	0.73	0.46	0.23	6.16
[redacted]	0.77	0.32	0.64	0.82	0.75	0.49	0.25	6.47
[redacted]	0.83	0.41	0.73	0.87	0.81	0.58	0.33	7.62
[redacted]	0.84	0.44	0.75	0.88	0.83	0.61	0.35	7.94
[redacted]	0.85	0.44	0.75	0.88	0.83	0.61	0.36	7.97
[redacted]	0.87	0.48	0.78	0.90	0.85	0.65	0.40	8.47
[redacted]	0.87	0.50	0.79	0.90	0.86	0.66	0.41	8.62
[redacted]	0.87	0.50	0.79	0.90	0.86	0.67	0.41	8.68
[redacted]	0.88	0.51	0.80	0.91	0.87	0.67	0.42	8.79
[redacted]	0.88	0.52	0.81	0.91	0.87	0.68	0.43	8.91
[redacted]	0.89	0.53	0.81	0.91	0.88	0.69	0.44	9.03
[redacted]	0.89	0.53	0.81	0.91	0.88	0.69	0.45	9.05
[redacted]	0.89	0.54	0.82	0.92	0.88	0.70	0.46	9.20
[redacted]	0.89	0.55	0.82	0.92	0.88	0.71	0.46	9.23
[redacted]	0.90	0.57	0.83	0.92	0.89	0.72	0.48	9.45
[redacted]	0.91	0.58	0.84	0.93	0.90	0.73	0.49	9.59
[redacted]	0.91	0.59	0.84	0.93	0.90	0.74	0.50	9.68
[redacted]	0.91	0.60	0.85	0.93	0.90	0.75	0.51	9.79
[redacted]	0.92	0.61	0.86	0.94	0.91	0.76	0.53	9.98
[redacted]	0.92	0.63	0.87	0.94	0.91	0.77	0.54	10.14
[redacted]	0.94	0.69	0.89	0.95	0.93	0.81	0.61	10.81
[redacted]	0.94	0.69	0.90	0.95	0.93	0.82	0.61	10.88
[redacted]	0.95	0.72	0.91	0.96	0.94	0.83	0.64	11.15
[redacted]	0.95	0.73	0.91	0.96	0.94	0.84	0.65	11.29
[redacted]	0.96	0.76	0.92	0.97	0.95	0.86	0.69	11.63
[redacted]	0.96	0.78	0.93	0.97	0.96	0.87	0.71	11.89
[redacted]	0.96	0.79	0.93	0.97	0.96	0.88	0.72	12.01
[redacted]	0.97	0.82	0.95	0.98	0.97	0.90	0.76	12.42
[redacted]	0.99	0.91	0.98	0.99	0.99	0.95	0.88	13.66

Instrument C-7: Evaluation Survey

Panelist Identification Number: ____

Please provide your unique identification number on the line above. Please note that your identification number is used for analysis purposes only. Your responses to these questions will be held in strict confidence and will be analyzed in conjunction with those of the other panelists who participated in this meeting. This evaluation form contains two parts. The first part asks for feedback on aspects of the workshop. The second part asks for your understanding, thoughts, feelings and actions during the workshop. Please address the questions in each of the parts as completely as possible. Your feedback will help in the improvement of similar workshops in the future.

PART 1: About the Workshop.

The questions in this part seek for your feedback on aspects of the workshop (Some examples of aspects of the workshop include: Orientation, review of test materials, PLD discussion, Instructions on the modified Angoff process, table discussions, whole group discussions, practice and real exercises, feedback, breaks and timing of activities, etc.)

1. What did you find most helpful about the training?

2. During the training what would you have liked to know more about?

3. Did you have questions or concerns that were not answered or addressed in the training session? Please indicate these below.

4. Please use the space below to provide additional comments concerning the adequacy, appropriateness, usefulness, or organization of the training.

Item no.	**Statement**	**Poor**	**Fair**	**Good**	**Very good**
5.	What is your overall assessment of performance of the workshop facilitators				
6.	What is your overall assessment of the training				

Instrument C-7: Evaluation Survey (cont'd)

PART II: About You as a Participant

Please indicate the extent to which each factor impacted your recommendations

Item no.	**Statement**	**Strongly Disagree**	**Disagree**	**Agree**	**Strongly Agree**
7.	What is measured by the items				
8.	Performance level descriptors				
9.	Your perception of the quality of items				
10.	Your educational or classroom experience				
11.	Discussions and Feedback				

Please indicate how confident you are about your recommendations

Item no.	**Statement**	**Strongly Disagree**	**Disagree**	**Agree**	**Strongly Agree**
12.	I am confident that I have a reasonable idea of the barely proficient students.				
13.	I am confident about my cut score recommendation.				

Please indicate your perception of your understanding of the standard setting process

Item no.	**Statement**	**Strongly Disagree**	**Disagree**	**Agree**	**Strongly Agree**
14.	I was able to follow instructions and complete the required ratings accurately.				
15.	I understood the tasks and feedback that were provided.				

Instrument C-7: Evaluation Survey (cont'd)

Please choose the rating that best reflects your thoughts, feelings and actions during the workshop. During the workshop

Item no.	**Statement**	**Strongly Disagree**	**Disagree**	**Agree**	**Strongly Agree**
16.	I was always tuned in.				
17.	I took active part in discussions.				
18.	I connected new material with what I already knew.				
19.	I talked with fellow participants about the training material.				
20.	When reflecting on the training material, I made connections with other things that I know.				

Please choose the rating that best reflects your thoughts, feelings and actions during the workshop. During the workshop

Item no.	**Statement**	**Strongly Disagree**	**Disagree**	**Agree**	**Strongly Agree**
21.	I adjusted my rating to incorporate ideas from all group members.				
22.	I adjusted my rating to incorporate ideas of the group members with the most convincing arguments.				
23.	I did not adjust my rating because I am confident about them.				
24.	I did not adjust my rating because I did not learn from the discussion.				
25.	I did not adjust my rating because I didn't want to use others' ideas				
26.	I enjoyed what I was doing.				
27.	I felt the activities were important to me.				
28.	I wish I had being doing something else.				

Appendix D:

Row Dimension Principal Coordinates Plots

Figure Appendix D-1: Heuristic Training Plot of PCO With Participant Table Group Point Labels

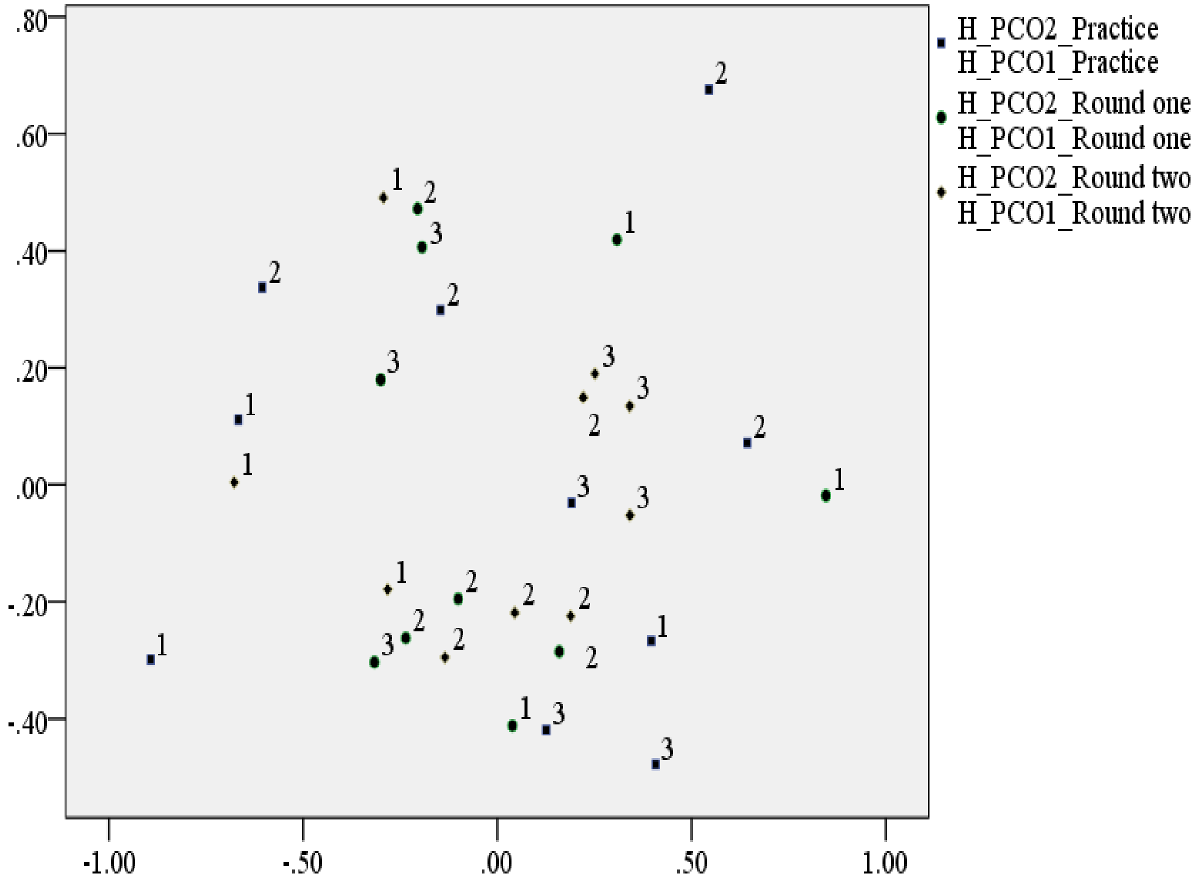


Figure Appendix D-2: Normative Training Plot of PCO With Participant Table Group Point Labels

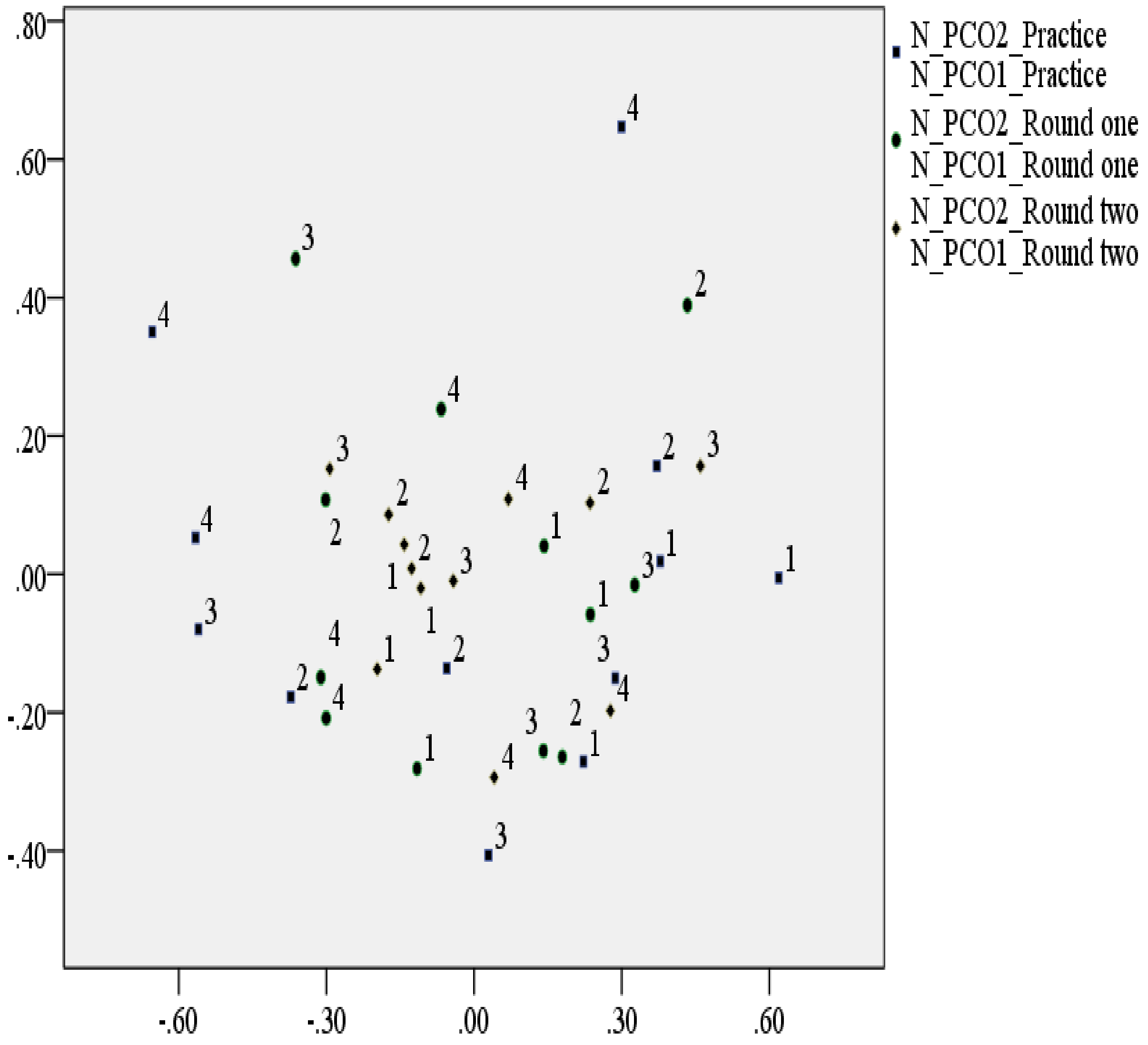


Figure Appendix D-3: Heuristic Training Plot of PCO With Indicator of Math Specialization

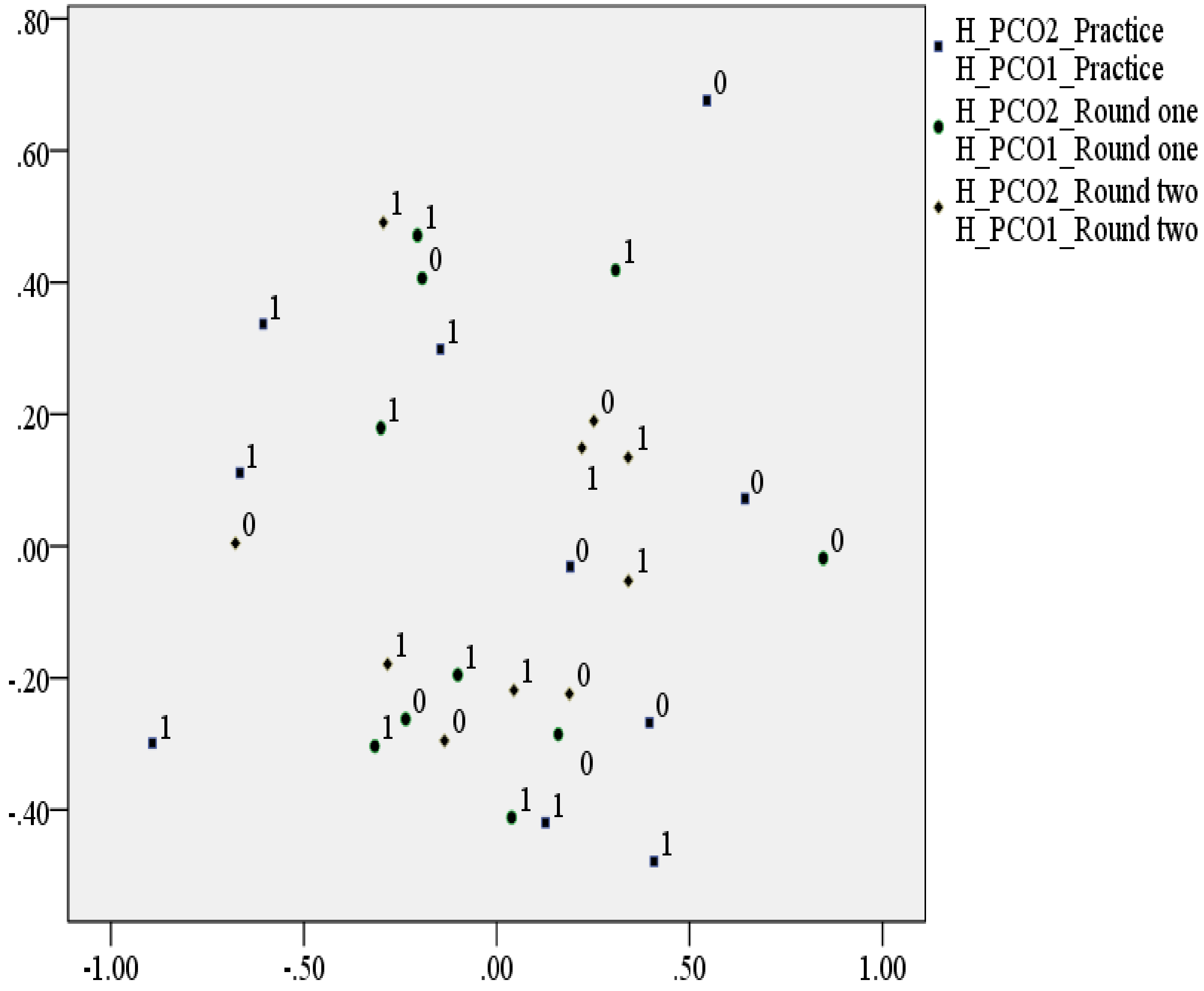


Figure Appendix D-4: Normative Training Plot of PCO With Indicator of Math Specialization

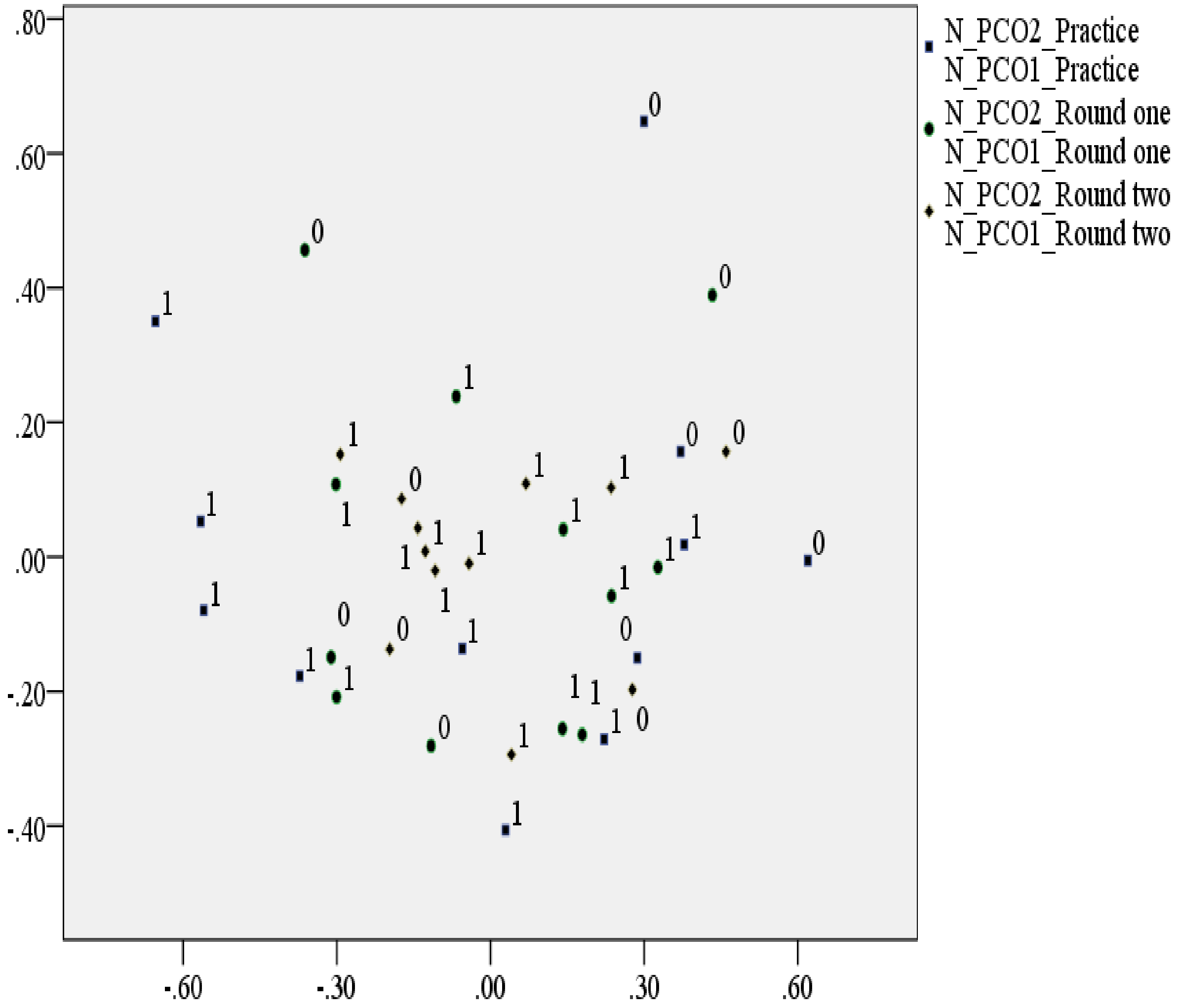


Figure Appendix D-5: Plot of Heuristic Training PCO With Indicator of Teaching Experience

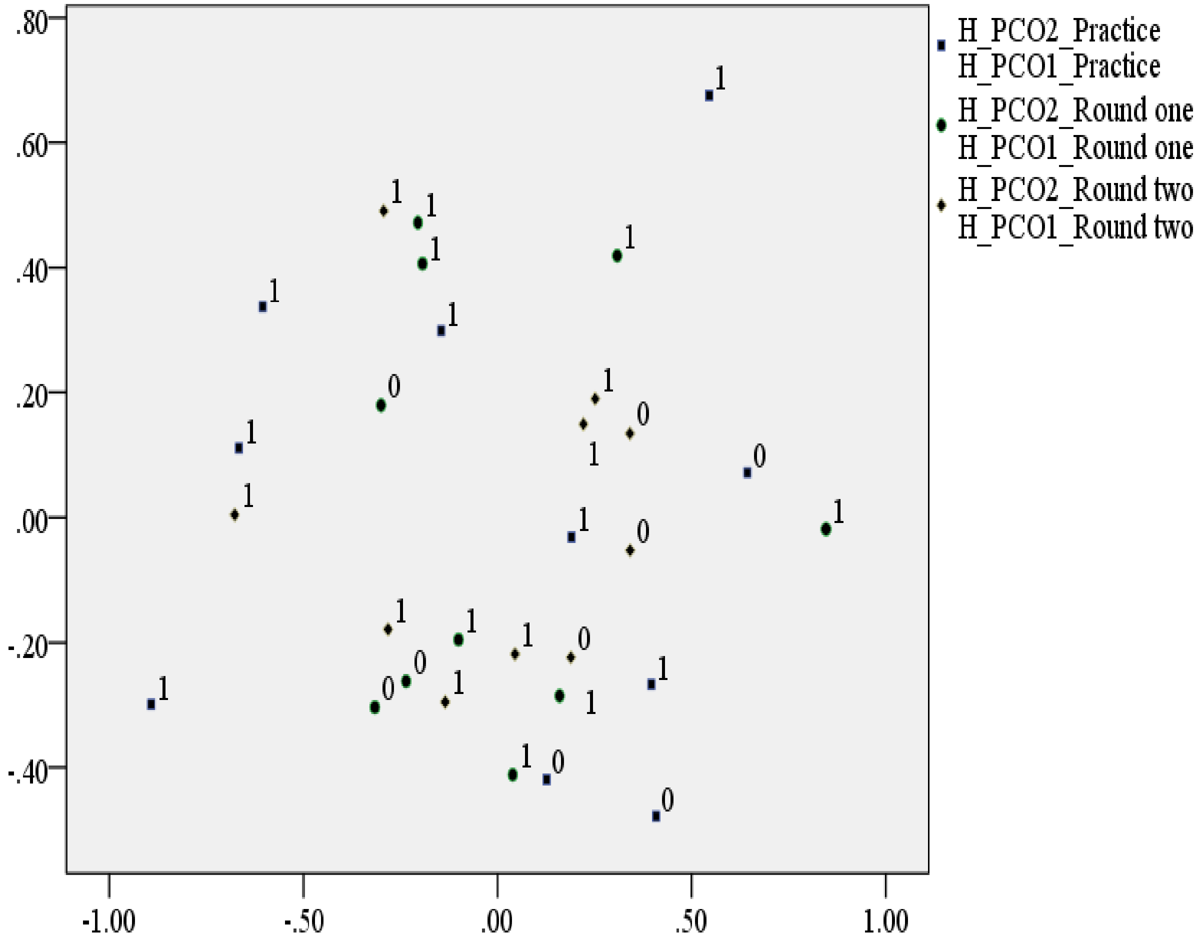
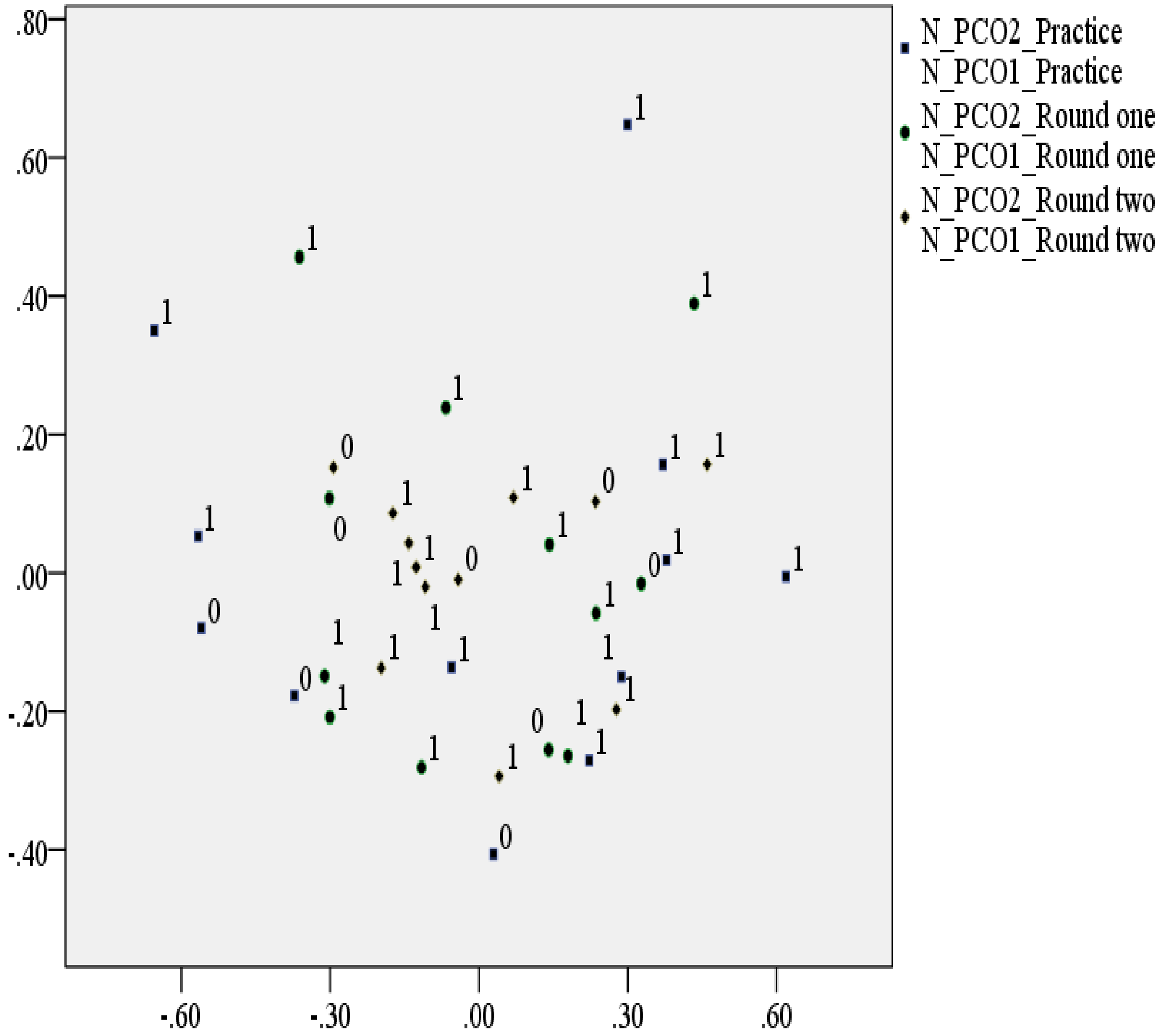


Figure Appendix D-6: Plot of Normative Training PCO With Indicator of Teaching Experience



REFERENCES

REFERENCES

Addis, D.R., Wong, A. T., & Schacter, D.L. (2007). Remembering the past and imagining the future: Common and distinct neural substrates during event construction and elaboration. *Neuropsychologica, 45,* 1363-1377.

Allport, G.W. (1954). *The nature of prejudice*. Reading, MA: Addison Wesley.

American College Testing (1995a, June). *Results of the 1994 geography NAEP achievement levels-setting pilot study July 14-18, 1994.* Iowa City, IA: Author.

American College Testing, (1995b, October). *Results of the 1994 U.S. history NAEP achievement levels-setting pilot study August 11-15, 1994*. Iowa City, IA: Author.

American College Testing (2005, April). *Developing achievement levels on the 2005 National Assessment of Educational Progress in grade twelve mathematics: Technical report*. Iowa City, IA: Author.

Anderson, J.R. (1990). *The adaptive character of thought*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Angoff, W.H. (1971). Scales, norms, and equivalent scores. In R.L. Thorndike (Ed.). *Educational measurement* (pp. 508-600). Washington, DC: American Council on Education.

Appleton, J.J., Christenson, S.L., Kim, D., & Reschly, A.L. (2006). Measuring cognitive and psychological engagement: Validation of the student engagement instrument. *Journal of School Psychology*, *44,* 427-45.

Arkes, H.R., & Hammond, K.R. (Eds). (1986) *Judgment and decision making.* Cambridge, MA: Cambridge University Press.

Atkinson, R.C., & Shriffin, R.M. (1968). Human memory: A Proposed system and its control processes. In K.W. Spence, & J.T. Spence (Eds.), *The Psychology of Learning and Motivation, (*Vol. 2). London: Academic Press.

Baddeley, A. (1992). Working memory. *Science*, *255,* 556-559.

Baddeley, A.D. & Hitch, G. (1974). Working memory. In G.H. Bower (Ed.), *The Psychology of Learning and Motivation,* (Vol. 8.). London: Academic Press.

Baron, J. (2000). *Thinking and deciding* (3rd ed.). New York: Cambridge University Press.

Baron, J. (2004). Normative models of judgment and decision making. In D. J. Koehler & N. Harvey (Eds.), *Blackwell handbook of judgment and decision making* (pp. 62-88). Oxford, UK: Blackwell.

Bartlett, F.C. (1932). *Remembering*. Cambridge University Press.

Beach, L. R., & Braun, G. P. (1994). Laboratory studies of subjective probability: A status report. In G. Wright & P. Avion (Eds.), *Subjective probability* (pp. 107-127). New York: Wiley.

Bejar, I. I. (1983). Subject matter experts' assessment of item statistics. *Applied Psychological Measurement, 7,* 303-310.

Berk, R. A. (1986). A consumer's guide to setting performance standards on criterion-referenced tests. *Review of Educational Research, 56*, 137–172.

Berstein, D.A., & Nash, P.W. (1999).*Essentials of psychology*. Houghton Mifflin Company.

Bolger F. & Wright, G. (1993). Coherence and calibration in expert probability judgment. *OMEGA International Journal of Management Science*, *21*(6), 629-644.

Bolger, F., & Wright, G. (1994). Assessing the quality of expert judgment: Issues and analysis. *Decision Support Systems* in Press.

Borg, I., & Groenen, P. (1997). *Modern multidimensional scaling: Theory and applications*. New York: Springer-Verlag.

Bower, G.H., Clark, M.C., Lesgold, A. M., & Winzenz, D. (1969). Hierarchical retrieval schemes in recall of categorized word lists. *Journal of Verbal Learning and Verbal Behavior, 8*, 323-343.

Brainerd, C.J., & Mojardin, A.H. (1998). Children's spontaneous false memories for narrative statements: Long-term persistence and mere-testing effects. *Child Development, 69,* 1361-1377.

Brandon, P.R. (2004). Conclusions about frequently studied modified Angoff standard-setting topics. *Applied Measurement in Education, 17*, 59-88.

Bransford, J.D., & Johnson, M.K. (1972). Contextual prerequisites for understanding: Some investigations of comprehension and recall. *Journal of Verbal Learning and Verbal Behavior, 11,* 717-726.

Broadbent, D.E. (1958). *Perception and communication.* Oxford: Pergamon.

Brown, J. (1976). *Recall and recognition.* John Wiley & Sons Ltd.

Brown, R., & Kulik, J. (1977). Flashbulb memories. *Cognition, 5*, 73-99.

Brunswik, E. (1955). Representative design and probabilistic theory in a functional psychology. *Psychological Review, 62,* 193-217.

Buckendahl, C.W. (2005). Qualitative inquiries of participants' experiences with standard setting. *Applied Measurement in Education, 18,* 219-221.

Busch, J.C., & Jaeger, R.M. (1990). Influence of type of judge, normative information, and discussion on standards recommended for the National Teacher Examinations. *Journal of Educational Measurement, 27,* 145-163.

Caines, J., & Engelhard, G. (2009, April). *Evaluating body of work judgments of standard-setting panelists*. Paper presented at the annual meeting of the American Educational Research Association. San Diego, CA.

Cammilli, G., Cizek, G.J., & Lugg, C.A. (2001). Psychometric theory and the validation of performance standards: History and future perspectives. In C. J. Cizek (Ed.), *Setting performance standards: Concepts, methods, and perspectives* (pp. 455-475). Mahwah, NJ: Lawrence Erlbaum.

Campbell (1971). Personnel training and development. *Annual Review Psychology, 22,*565-602.

Carroll, J. D., & Wish, M. (1974). Multidimensional perceptual models and measurement methods. In E. C. Carterette & M. P. Friedman (Eds.), *Handbook of perception.* New York: Academic Press.

Chiang, H. (2009). How accountability pressure on failing schools affects student achievement. *Journal of Public Economics*, 93, 1045–1057.

Cizek, G.J. (1991). *An investigation into one alternative to the group-process procedure for setting performance standards on a medical specialty examination*. (Unpublished doctoral dissertation). Michigan State University, East Lansing, MI.

Cizek, G. J. (1993). Reconsidering standards and criteria. *Journal of Educational Measurement, 30*, 93–106.

Cizek, G. J. (2001). Conjectures on the rise and call of standard setting: An introduction to context and practice. In C. J. Cizek (Ed.), *Setting performance standards: Concepts, methods, and perspectives* (pp. 3-17).Mahwah, NJ: Lawrence Erlbaum.

Cizek, G. J., & Bunch, M. B. (2007). *A guide to establishing and evaluating performance standards on tests*. Thousand Oaks, CA: Sage Publications.

Clauser, B. E., Harik, P., Margolis, M. J., McManus, I. C., Mollon, J., Chis, L., & Williams, S.(2009a). An empirical examination of the impact of group discussion and examinee performance information on judgments made in the Angoff standard-setting procedure, *Applied Measurement in Education*, *22*(1), 1 – 21.

Clauser, B.E., Mee, J., Baldwin, S.G., Margolis, M.J., & Dillon, G.F. (2009b). Judges' use of examinee performance data in an Angoff standard-setting exercise for a medical licensing examination: An experimental study. *Journal of Educational Measurement, 46,* 390-407.

Clauser, B. E., Mee, J., & Margolis, M. J. (2011, April). *The effect of data format on integration of performance data into Angoff judgments*. Paper presented at the Annual Meeting of the National Council on Measurement in Education, New Orleans, LA.

Clauser, B.E., Swanson, D.B., & Harik, P. (2002). A multivariate generalizability analysis of the impact of training and examinee performance information on judgments made in an Angoff-style standard-setting procedure. *Journal of Educational Measurement, 39,* 269-290.

Collins, A.M., & Loftus, E. (1975). A spreading activation theory of semantic memory. *Psychological Review, 82,* 407-428.

Collins, A.M., & Quillian, M.R. (1969). Retrieval time from semantic memory. *Journal of Verbal Learning & Verbal Behavior, 8,* 240-248.

Cooper, M. (1995). *Comparison of three methods of assessing teacher's judgmental accuracy*. (Unpublished master's thesis). University of Nebraska, Lincoln.

Craik, F.I.M., & Lockhart, R.S. (1972). Levels of processing: A framework for memory research. *Journal of Verbal Learning and Verbal Behavior, 11,* 671-684.

Craik, F.I.M., & Tulving, E. (1975) Depth of processing and retention of words in episodic memory. *Journal of Experimental Psychology: General, 104*(3), 268-294.

Cross, L., Impara, J., Frary, R., & Jaeger, R. (1984). A comparison of three methods for establishing minimum standards on the National Teacher Examinations. *Journal of Educational Measurement, 21,* 113-130.

Diana, R.A., Yonelinas, A.P., & Ranganath, C. (2007). Imagining collection and familiarity in the medial temporal lobe: A three-component model. *Trends in Cognitive Sciences, 11,* 379-386.

Dillon S. (2011, August 8). Retrieved from http://www.nytimes.com/2011/08/08/education/08educ.html?_r=3&partner.

Dingle, H. (1950). A theory of measurement. *The British Journal for the Philosophy of Science, 1*(1)*,* 5-26.

Dion K., Berscheid E., & Walster E. (1972). What is beautiful is good. *Journal of Personality and Social Psychology, 24,* 285–90.

Druckman, D., & Bjork, R.A. (1994). *Learning, remembering, believing: Enhancing human performance*. Washington, DC: National Academy Press.

Efron, B. (1979). Bootstrap methods: Another look at the jackknife, *The Annals of Statistics*, *7,* 1-26.

Engelhard, G. (2007). Evaluating Bookmark judgments. *Rasch Measurement Transactions*, *21*, 1097-1098.

Engelhard, G. (in press). Evaluating the judgments of standard-setting panelists using Rasch measurement theory. In E. V. Smith, Jr., & G. E. Stone (Eds.), *Applications of Rasch measurement in criterion-referenced testing,* JAM Press.

Engelhard, G., & Anderson, D. W. (1998). A binomial trials model for examining the ratings of standard-setting judges. *Applied Measurement in Education, 11,* 209-230.

Engelhard, G., & Stone, G.E. (1998). Evaluating the quality of ratings obtained from standard-setting judges. *Educational and Psychological Measurement, 58,* 179-196.

Evans, J.S. (2008). Dual-processing accounts of reasoning, judgment, and social cognition. *Annual Review of Psychology, 59*, 255–278.

Eysenck, M.W. (1979). Depth, elaboration, and distinctiveness. In L.S.Cermak & F.I.M. Craik (Eds.), *Levels of processing in human memory*. Hillsdale, NJ: Lawrence Erlbaum Associates Inc.

Eysenck, M.W., & Eysenck, M.C. (1980). Effects of processing depth, distinctiveness, and word frequency on retention. *British Journal of Psychology, 71,* 263-274.

Eysenck, M.W., & Keane, M.T. (2010). *Cognitive psychology: A student's Handbook* (6th ed.). Hove and New York, Psychology Press, Taylor and Francis group.

Ferdous, A.A., & Plake, B.S. (2005). Understanding the factors that influence decisions of panelists in standard setting-study. *Applied Measurement in Education*, *18,* 223-232.

Festinger, L. (1954). A theory of social comparison processes. *Human Relations, 7*, 114-140.

Figlio, D. N., & Rouse, C.E. (2006). Do accountability and voucher threats improve low-performing schools? *Journal of Public Economics, 90*, 239– 255.

Fitzpatrick, A.R. (1989). Social influences in standard-setting: The effects of social interaction on group judgments. *Review of Educational Research*, *59*, 315-328.

Fryer, R.G., & Jackson, M.O. (2007, November). *A categorical model of cognition and biased decision-making*. Unpublished manuscript.

Fuhrman, S.H., & Elmore, R.F. (2004). *Redesigning accountability systems for education*. Teachers college, Columbia University New York London.

Gardiner, J.M., & Java, R.I. (1993). Recognizing and remembering. In A.F Collins, S.E. Gathercole, M.A. Conway, & P.E. Morris, *Theories of memory* (Eds.). UK: Lawrence Erlbaum Associates Hove.

Gati, I., & Tversky, A. (1984). Weighting common and distinctive features in perceptual and conceptual judgments. *Cognitive Psychology, 16,* 341-370.

Gigerenzer, G. (1994). Why the distinction between single-event probabilities and frequencies is relevant for psychology and vice versa. In G. Wright & P. Avion (Eds.), *Subjective probability* (pp. 129-162). New York: Wiley.

Gigerenzer, G. (1996). On narrow norms and vague heuristics: A reply to Kahneman and Tversky. *Psychological Review, 103*(3), 592-596.

Gigerenzer, G. (2008). *Rationality for mortals: How people cope with uncertainty*. New York: Oxford University Press.

Gigerenzer G., & Gaissmaier, W. (2011). Heuristic decision making. *The Annual Review of Psychology, 62,* 451-482.

Gigerenzer, G., & Hoffrage, U. (1995). How to improve Bayesian reasoning without instruction: Frequency formats. *Psychological Review*, *102*, 684-704.

Gigerenzer, G., & Murray, D.J. (1987). *Cognition as intuitive statistics.* Hillsdale, NJ: Erlbaum.

Gigerenzer, G., Todd, P. M., & the ABC Research Group (1999). *Simple heuristics that make us* smart. New York: Oxford University Press.

Giraud, G., & Impara, J.C. (2005). Making the cut: The cut score setting process in a public school district. *Applied Measurement in Education, 18,* 289-312.

Giraud, G., Impara, J.C., & Plake, B.S. (2005). Teachers' conceptions of target examinee in Angoff standard setting. *Applied Measurement in Education, 18,* 223-232.

Glass, G.V. (1978). Standards and criteria. *Journal of Educational Measurement, 15,* 237-261.

Glenberg, A.M., Smith, S.M., & Green, C. (1977). Type I rehearsal: Maintenance and more. *Journal of Verbal Learning & Verbal Behavior, 16,* 339-352.

Godden, D.R., & Baddeley, A.D. (1975). Context dependent memory in two natural environments: On land and underwater. *British Journal of Psychology, 66,* 325-331.

Goldstein (1980). Training in work organizations. *Annual Review Psychology, 31*, 229-72.

Goldstein, E.B. (2008). *Cognitive psychology: Connecting mind, research, and everyday experience* (2nd ed.). Wadsworth, Cengage Learning.

Good, P. (2006). *Resampling methods* (3rd ed.). Birkhauser.

Graf, P., & Schacter, D.L. (1985). Implicit and explicit memory for new associations in normal and amnesic subjects. *Journal of Experimental Psychology: Learning, Memory & Cognition, 11,* 501-518.

Haertel, E.H., & Lorie, W.A. (2004). Validating standards-based test score interpretations. *Measurement: Interdisciplinary Research and Perspectives, 2*(2), 61-103.

Halpin, G., & Halpin, G. (1983). *Reliability and validity of 10 different standard setting procedures*. Paper presented at the American Psychological Association, Anaheim.

Hambleton, R.K., & Pitoniak, M.J. (2006). Setting performance standards. In R.L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 433-470). Westport, CT: American Council on Education/ Praeger.

Hand, D.J. (1996). Statistics and the theory of measurement. *Journal of Royal Statistical Society A.159*(3), 445-492.

Hanick, P. L. (1998a, September). *Civics pilot study evaluation questionnaire summary report.* Report to the Technical Advisory Committee on the Standard Setting, Minneapolis, MN.

Hanick, P. L. (1998b, October). *Writing pilot study process evaluation questionnaire summary report.* Report to the Technical Advisory Committee on the Standard Setting, Detroit, MI.

Hanick, P. L. (1999, February). *1998 Civics NAEP achievement levels-settings meeting: Summary report of process evaluation questionnaires*. Report to the Technical Advisory Committee on the standard setting. Atlanta, GA.

Harris, R.J., Sardarpoor-Bascom, F., & Meyer, T. (1989). The role of cultural knowledge in distorting recall for stories. *Bulletin of the Psychonomic Society, 27,* 9-10.

Harvey, N. (2001). Studying judgment: General issues. *Thinking and Reasoning, 7,* 103-118.

Hassabis, D., Kumaran, D. Van, S.D., & Maguire, E.A. (2007). Patients with hippocampal amnesia cannot imagine new experiences. *Proceedings of the National Academy of Sciences of the United States of America, 104*, 1726-1731.

Hastie, R. (2001). Problems for judgment and decision making. *Annual Review of Psychology, 52,* 653-683.

Hays, W. L. (1994). *Statistics* (5th ed.). Fort Worth: Harcourt Brace College Publishers.

Hein, S.F., & Skaggs, G.E. (2009). A qualitative investigation of panelists' experiences of standard setting using two variations of the Bookmark Method. *Applied Measurement in Education. 22*(3) 207-228.

Hein, S.F., & Skaggs, G.E. (2010). Conceptualizing the classroom of target students: A qualitative investigation of panelists' experiences during standard setting. *Educational Measurement: Issues and Practice, 22*(2), 36 – 44.

Hoge, R. D., & Coladarci, T., (1989). Teacher-based judgments of academic achievement: A review of the literature. *Review of Educational Research*, *59*(3), 297-313.

Howell, D. C. (2002). *Statistical methods for psychology* (5th ed.). Pacific Grove, CA: Wadsworth Publishing.

Huck, S.W. (2004). *Reading statistics and research* (4th ed.). Pearson Education Inc.

Impara, J. C., & Plake, B. S. (1997). Standard setting: An alternative approach. *Journal of Educational Measurement, 34*, 353–366.

Impara, J. C., & Plake, B. S. (1998). Teachers' ability to estimate item difficulty: A test of the assumptions of the Angoff standard setting method. *Journal of Educational Measurement*, *35*, 69–81.

Jacob, B.A. (2005). Accountability, incentives and behavior: the impact of high-stakes testing in the Chicago Public Schools. *Journal of Public Economics, 89*, 761–796.

Jacoby, L.L. (1983). Remembering the data: Analyzing interactive process in reading. *Journal of Verbal Learning and Verbal Behavior, 22,* 485-508.

Jacoby, L.L. (1988). Memory observed and memory unobserved. In U. Neisser & E. Winograd (Eds.), *Remembering reconsidered: ecological and traditional approaches to the study of memory* (pp. 145-177). New York: Cambridge University Press.

Jacoby, L.L., Kelly, C.M., & Dywan, J. (1989). Memory attributions. In H.L. Roediger III & F.I.M. Craik (Eds.), *Varieties of memory and consciousness: Essays in honor of Endel Tulving* (pp. 391-422). Hillsdale, NJ: Lawrence Erlbaum Associates Inc.

Jaeger, R.M. (1989). Certification of student competence. In R.L. Linn (Ed.), *Educational Measurement* (3rd ed., pp. 485-514). Washington, DC: American Council on Education.

Jaeger, R.M. (1991). Selection of judges for standard-setting. *Educational Measurement: Issues and Practice, 10*(2), 3-6.

Jenkins, J. J. (1979). Four points to remember: A tetrahedral model of memory experiments. In L.S. Cermak & F.I.M. Craik (Eds.), *Levels of processing in human memory* (pp. 429-446). Hillsdale, NJ: Erlbaum.

Jenkins, J.J., & Russell, W.A. (1952). Associative clustering during recall. *Journal of Abnormal and Social Psychology, 47,* 818-821.

Johnson, E.J., & Tversky, A. (1983). Affect, generalization, and the perception of risk. *Journal of Personality and Social Psycholology, 45,* 20–31.

Kahneman, D., & Tversky, A. (1972). Subjective probability: A judgment of representativeness. *Cognitive Psychology*, *3*(3), 430-454.

Kahneman, D. & Tversky, A. (1982). On the study of statistical intuitions. *Cognition*, *11*, 123-141.

Kane, M. T. (1987). On the use of IRT models with judgmental standard-setting procedures. *Journal of Educational Measurement*, *24*, 333–345.

Kane, M. T. (1994). Validating the performance standards associated with passing scores. *Review of Educational Research*, *64*, 425-461.

Kane, M.T. (2001). So much remains the same: Conception and status of validation in setting standards. In C. J. Cizek (Ed.), *Setting performance standards: Concepts, methods, and perspectives* (pp. 53–88).Mahwah, NJ: Lawrence Erlbaum.

Karantonis, A., & Sireci, S. G. (2006). The Bookmark standard-setting method: A literature review. *Educational Measurement: Issues and Practice, 21*(1), 4-12.

Keppel, G., & Underwood, B.J. (1962). Proactive inhibition in short-term retention of single items. *Journal of Verbal Learning and Verbal Behavior, 1,* 153-161.

Keren G. (1991). Calibration and probability judgments: Conceptual and methodological issues. *Acta Psychologica, 77*, 217-273.

Keren, G., & Teigen, K.H. (2004). Yet another look at the Heuristics and Biases approach. In D. J. Koehler & N. Harvey (Eds.), *Blackwell handbook of judgment and decision making* (pp. 62-88). Oxford, UK: Blackwell.

Kerlinger, F. N. (1986). *Foundations of behavioral research* (3rd ed.). New York: Holt, Rinehart and Winston.

Kintsch, W. (1970). Models for free recall and recognition. In D.A. Norman (Ed.), *Models of Human Memory*. London: Academic Press.

Kirkpatrick, D. (1994). *Evaluating training programs*: *The four levels.* San Francisco, CA: Berrett-Keohler.

Kirkpatrick, D.L., & Kirkpatrick, J.D. (2006). *Evaluating training programs: The four levels,* (3rd ed.). Berrett-Koehler Publishers, Inc. San Francisco.

Koriat, A., & Goldsmith, M. (1996). Memory metaphors and real-life/laboratory controversy: Correspondence versus storehouse conceptions of memory. *Behavioral & Brain Sciences, 19,* 167-188.

Lamm, H., & Myers, D.G. (1978). Group-induced polarization of attitudes and behavior. In L. Berkowitz (Ed.), *Advances in experimental social psychology* (Vol. 11). New York: Academic Press.

Larrick, R.P. (2004). Debiasing. In D. J. Koehler & N. Harvey (Eds.), *Blackwell handbook of judgment and decision making* (pp. 62-88). Oxford, UK: Blackwell.

Latham, G.P. (1988). Human resource training and development. *Annual Review Psychology, 39,* 545-82.

Lerner, B. (1979). Tests and standards today: Attacks, counterattacks, and responses. In R.T. Lennon (Ed.), *New directions for testing and measurement: Impactive changes on measurement* (pp. 15-31). San Francisco: Jossey Bass.

Lerner, J.S., Gonzalez, R.M., Small, D.A., & Fischoff, B. (2005). Effects of fear and anger on perceived risks of terrorism: A national field experiment. In S. Wessely & V.N. Krasnov (Eds.), *Psychological responses to the new terrorism: A NATO-Russia dialogue.* Amsterdam: IOS Press.

Lichtenstein, S., Slovic, P., Fischhoff, B., Layman, M., & Combs B. (1978). Judged frequency of lethal events. *Journal of Experimental Psychology Human Learning and Memory, 4,* 551–78.

Livingston, S. A., & Zieky, M. J. (1982). *Passing scores: A manual for setting standards of performance on educational and occupational tests.* Princeton, NJ: Educational Testing Service.

Loewenstein, G.F., Weber, E., Hsee, C.K., & Welch, N. (2001). Risk as feelings. *Psychological Bulletin 127,* 267–286.

Loomis, S. C. (2000, April). *Research study of the 1998 civics NAEP achievement levels*. Paper presented at the Annual Meeting of the American Educational Research Association, New Orleans, LA.

Lorge, I., & Kruglov, L. K. (1953). The improvement of the estimates of test difficulty. *Educational and Psychological Measurement, 13,* 34-46.

Mandler, G. (1980). Recognizing – the judgment of previous occurrence. *Psychological Review, 87,* 252-271.

Mantel, N. (1967). The detection of disease clustering and a generalized regression approach. *Cancer Research 27* (2): 209–220.

Mantyla, T. (1986). Optimizing cue effectiveness: Recall of 500 and 600 incidentally learned words. *Journal of Experimental Psychology: Learning, Memory and Cognition, 12,* 66-71.

Margolis, M. J. (2011, April). *The impact of examinee performance data on Angoff study Outcomes*. Paper presented at the Annual Meeting of the National Council on Measurement in Education, New Orleans, LA.

Mari, L. (2005). The problem of foundations of measurement. *Measurement,* 38.

McCloskey, M.E., & Glucksberg, S. (1978). Natural categories: Well defined or fuzzy sets? *Memory & Cognition, 6,* 462-472.

McCormick, J. (1987, August 17). The wisdom of Solomon. *Newsweek*, 24-25.

McGinty, D. (2005). Illuminating the "Black Box" of standard setting: An exploratory qualitative study. *Applied Measurement in Education*, *18*(3), 269–287.

Medin, D.L., Goldstone, R.L., & Gentner, D. (1993). Respects for similarity. *Psychological Review 100*(2) 254-278.

Mee, J., Clauser, B. E., & Margolis, M. J. (2011, April). *The impact of process instructions on judges' use of examinee performance data*. Paper presented at the Annual Meeting of the National Council on Measurement in Education, New Orleans, LA.

Messick, S. (1989). Validity. In R. L. Linn (Ed), *Educational Measurement* (3rd ed., pp. 13-103). New York: American Council on Education and Macmillan.

Michell, J. (1990). *An introduction to the logic of Psychological Measurement*. Lawrence Erlbaum Associates, Inc.

Michell , J. (1999). *Measurement in Psychology*. Cambride University Press.

Miller, G.A. (1956). The magical number seven, plus or minus two. Some limits on our capacity for processing information. *Psychological Review, 63,* 81-97.

Mills, C.N., Melican, G. J., Ahluwalia, N.T. (1991). Defining minimal competence. *Educational Measurement: Issues and Practice, 10*(2), 15-16.

Mitzel, H. C., Lewis, D. M., Patz, R. J., & Green, D. R. (2001). The Bookmark procedure: psychological perspectives. In G. J. Cizek (Ed.), *Setting performance standards: Concepts, methods, and perspectives* (pp. 249–281). Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

Moe, T.M. (2003). Politics, Control, and the Future of School Accountability. In Peterson, P.E. & West M.R. (Ed.). *No Child left behind?: the politics and practice of school accountability*. The Brookings Institution Press Washington, D.C.

Mooney, C. Z., & Duval, R D (1993). *Bootstrapping. A nonparametric approach to statistical inference.* Sage university paper series on quantitative applications in the social sciences, 07-095. Newbury Park, CA: Sage.

Morris, C.D., Bransford, J.D., & Franks, J.J. (1977). Levels of processing versus transfer appropriate processing. *Journal of Verbal Learning and Verbal Behavior, 16,* 519-533.

Murdoch, B. B. (1962). The serial position effect in free recall. *Journal of Experimental Psychology, 64,* 482-488.

Murphy, G. L. (2002). *The big book of concepts*. Cambridge, MA: The MIT Press.

Murphy, G.L., & Medin, D.L. (1985). The role of theories in conceptual coherence. *Psychological Review, 92*, 289–316.

Myers, D. G. (2004). *Psychology* (7th ed.). New York, Worth Publishers.

Nichols, P., Twing, J., Mueller, C.D., & O'Malley, K. (2010). standard-setting methods as measurement processes. *Educational Measurement: Issues and Practice. 29*(1), 14–24.

Norcini, J.J., Shea, J.A., & Kanya, D.T. (1988). The effect of various factors on standard setting. *Journal of Educational Measurement, 25,* 57-65.

Nosofsky, R. M. (1988a). Exemplar-based accounts of relations between classification, recognition, and typicality. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 14,* 700-708.

Nosofsky, R. M. (1988b). Similarity, frequency, and category representations. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 14,* 54-65.

O'Day, J.A. (2002). Complexity, accountability, and school improvement. *Harvard Educational Review, 72*(2).

Over, D. (2004). Rationality and the normative/descriptive distinction. In D. J. Koehler & N. Harvey (Eds.), *Blackwell Handbook of judgment and decision making* (pp. 62-88). Oxford, UK: Blackwell.

Perie, M. (2008). A guide to understanding and developing performance-level descriptors. *Educational Measurement: Issues and Practice, 27*(4), 15-29.

Petersen, L.R., & Petersen, M.J. (1959). Short-term retention of individual verbal items. *Journal of Experimental Psychology, 58,* 193-198.

Peterson, P.E., & West, M.R. (2003). *No Child left behind?: the politics and practice of school accountability*. The Brookings Institution Press Washington, D.C.

Plake, B. S., Melican, G. J., & Mills, C. N. **(**1991). Factors influencing intrajudge consistency during standard setting. *Educational Measurement: Issues and Practice, 10*(2), 15-16.

Popper, K.R. (1959). The propensity interpretation of probability. *The British Journal for the Philosophy of Science, 10*(37), 25 – 42.

Quillian, M.R. (1967). Word concepts. A theory and simulation of some basic semantic capabilities. *Behavioral Science, 12,* 410-430.

Quillian, M.R. (1969). The teachable language comprehender: A simulation program and theory of language. *Communications of the ACM, 12,* 459-476.

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests.* Copenhagen: Danmarks Paedagogiske Institut.

Raykov, T. & Marcoulides, G.A. (2011). *Introduction to psychometric theory*. Routledge Taylor & Francis Group New York London.

Raymond, M. R., & Reid, J. B. (2001). Who made thee a judge? Selecting and training participants for standard setting. In C. J. Cizek (Ed.), *Setting performance standards: Concepts, methods, and perspectives* (pp. 119–157). Mahwah, NJ: Lawrence Erlbaum.

Reckase, M. D. (2000). *The evolution of the NAEP achievement levels setting process: A summary of research and development efforts conducted by NAEP.* Iowa City: ACT.

Reckase, M.D. (2001). Innovative methods for helping standard-setting participants to perform their task: The role of feedback regarding consistency, accuracy, and impact. In C. J. Cizek (Ed.), *Setting performance standards: Concepts, methods, and perspectives* (pp. 159-173). Mahwah, NJ: Lawrence Erlbaum.

Reckase , M.D. (2006). A conceptual framework for psychometric theory of standard setting with examples of its use for evaluating the functioning of two standard setting methods. *Educational measurement: Issues and Practice*, *25*(2), 4-18.

Reckase, M.D (2009). Standard setting theory and practice: issues and difficulties. In N. Figueras & J. Noijons (Eds.), *Linking to the CEFR levels: research perspectives* (pp. 13 - 20). Arnhem: Cito, Institute for Educational Measurement Council of Europe European Association for Language Testing and Assessment (EALTA).

Reichenbach, H. (1949): *The theory of probability*, University of California Press.

Reid, J. B. (1991). Training judges to generate standard setting data. *Educational Measurement: Issues and Practice*, *10*(2), 11–14.

Roediger, H. L. (1990). Implicit memory: Retention without remembering. *American Psychologist, 45,* 1043-1056.

Roediger, H.L. (2008). Relativity of remembering: Why the laws of memory vanished. *Annual Review of psychology*, *59,* 225-254.

Rosch, E. (1973). Natural categories. *Cognitive Psychology, 4*, 328-350.

Rosch, E., & Mervis, C. B. (1975). Family resemblances: Studies in the internal structure of categories. *Cognitive Psychology, 7,* 573-603.

Rosch, E.H. , Mervis, C.B., Gray, W.D., Johnson, D.M., & Boyes-Braem, P. (1976). Basic objects in natural categories. *Cognitive Psychology*, *8*, 382-439.

Rudalevige, A. (2003). No Child Left Behind: Forging a congressional compromise**.** In P.E. Peterson & M.R. West (Eds.), *No Child Left Behind?: The politics and practice of school accountability*(pp. 23-54). Washington, DC: The Brookings Institution.

Ryan, R.M., & E.L. Deci (2000). Intrinsic and extrinsic motivations: Classic definitions and new directions. *Contemporary Educational Psychology, 25,* 54-67.

Sachs, J. (1967). Recognition memory for syntactic and semantic aspects of a connected discourse. *Perception & Psychophysics, 2,* 437-442.

Salas, E. & Cannon-Bowers, J.A. (2001).The science of training: A decade of progress. *Annual Review of Psychology*, *52*, 471-99.

Schacter, D.L. (1987). Implicit memory. History and current status. *Journal of Experimental Psychology: Learning, Memory and Cognition, 13,* 501-518.

Schacter, D.L. (1990). Perceptual representation systems and implicit memory: Toward a resolution of the multiple memory systems debate. In A. Diamond (Ed.), Development and neural bases of higher cognition. *Annals of New York Academy of Sciences, 608*, 543-571.

Schacter, D.L., & Addis, D.R. (2007). The cognitive neuroscience of constructive memory: Remembering the past and imagining the future. *Philosophical Transactions of Royal Society, B, 362,* 773-786.

Schacter, D.L., & Tulving, E. (1994). What are the memory systems of 1994? In D.L. Schacter & E. Tulving (Eds.), *Memory systems*. Cambridge, MA: MIT Press.

Schaeffer, G. A, & Collins, J. L. (1984, April). *Setting performance standards for high-stakes tests*. Paper presented at the Annual Meeting of the National Council on Measurement in Education, New Orleans.

Schiffman, S.S., Reynolds, M. L., & Young, F.W. (1981). *Introduction to multidimensional scaling: Theory, methods and applications.* London, UK: Academic Press, Inc.

Schultz, E.M. (2006). Commentary: A response to Reckase's conceptual framework and examples for evaluating standard setting methods. *Educational Measurement: Issues and Practice, 25*(3), 4-13.

Schwarz, N., & Clore ,G.L. (1983). Mood, misattribution, and judgments of well-being: informative and directive functions of affective states. *Journal of Personality and Social Psychology, 45*, 513–23.

Scott, C., Duffrin, E., Kelleher, M., & Neuman-Sheldon B. (2009). Improving low- performing schools: Lessons from five years of studying school restructuring under No Child Left Behind. *Research Report of Center on Education Policy*. Washington, DC.

Shafir, E., & LeBoeuf, R.A. (2002). Rationality. *Annual Review of Psychology, 53*, 491-517.

Shepard, R. N. (1974). Representation of structure in similarity data: Problems and prospects. *Psychometrika, 39,* 373-421.

Shepard, L. A. (1994, October). *Implications for standard setting of the NAE evaluation of NAEP achievement levels.* Paper presented at the Joint Conference on Standard Setting for Large Scale Assessments, National Assessment Governing Board, National Center for Educational Statistics, Washington, DC.

Shepard, L. A. (1995). Implications for standard setting of the National Academy of Education evaluation of the National Assessment of Educational Progress achievement levels. In L. Crocker & M. Zieky (Eds.), *Proceedings of the Joint Conference on Standard Setting for Large-Scale Assessments,* Vol. 2 (pp. 143–160). Washington, DC: U.S. Government Printing Office.

Shepard, L., Glaser, R., Linn, R., & Bohrnstedt, G. (1993). *Setting standards for student achievement*. Stanford, CA: National academy of Education.

Shrock, A. S., & Coscarelli, W.C. (2007). *Criterion referenced test development: Technical and legal guidelines for corporate training* (3rd ed.). San Francisco, CA: John Wiley & Sons, Inc.

Simon, H. A. (1957). *Models of man: Social and rational*. New York: Wiley.

Simon, H. A. (1990). Invariants of human behavior. *Annual Review of Psychology, 41*, 1-19.

Skaggs, G., & Hein, S.F. (2011).Reducing the cognitive complexity associated with standard setting: A comparison of the Single-Passage Bookmark and Yes/No methods. *Educational and Psychological Measurement.*

Sloman, S. A., & Rips, L. J. (1998). Similarity as an explanatory construct. *Cognition, 65* , 87-101.

Slovic, P., & Lichtenstein, S. (1971). Comparison of Bayesian and regression approaches to the study of information processing in judgment. *Organizational Behavior and Human Performance, 6*(6), 649-744.

Smith, E. E., & Medin, D. L. (1981). *Categories and concepts.* Cambridge, MA: Harvard University Press.

Smith, E.E., Shoben, E.J., & Rips, L., (1974). Structure and process in semantic memory: A featural model for semantic decisions. *Psychological Review, 81*, 214–241.

Sperling, G. (1960). The information available in brief visual presentations. *Psychological Monographs: General and Applied, 74,* (11, whole No. 498), 1-29.

Spiers, H.J., Maguire, E.A., & Burgess, N. (2001). Hippocampal amnesia. *Neurocase*, *7*, 357-382.

Staresina, B.P., & Davachi, L. (2006). Differential encoding mechanisms for subsequent associative recognition and free recall. *Journal of Neuroscience, 26,* 9162-9172.

Stevens, S.S. (1946). On the theory of scales of measurement. *Science, 103*(2684), 677-680.

Stoner, J. A.F. (1961). *A comparision of individual and group decisions involving risk*. (Unpublished master's thesis), Massachusetts Institute of Technology, School of Industrial Management, Cambridge, MA.

Suen, H. K. (1990). *Principles of test theories*. Hillsdale, NJ: Lawrence Erlbaum.

Sun, W. (2003). *Interpretations of probability*. (Unpublished doctoral dissertation), University of Connecticut, Storrs, CT.

Tennenbaum, S.I., & Yukl G. (1992). Training and development in work organizations. *Annual Review Psychology, 43*, 399-441.

Thomdike, R. L. (1980). *Item and score conversion by pooled judgment.* Paper presented at the Educational Testing Service Conference on Test Equating, Princeton, NJ.

Todd, P. M., & Gigerenzer, G. (2007). Environments that make us smart: Ecological rationality. *Current Directions in Psychological Science, 16*(3), *167-171.*

Torre, M. , & Gwynne, J. (2009a). When schools close: Effects on displaced students in Chicago Public Schools. *Research report of consortium on Chicago school research at the University of Chicago Urban Education Institute.* Chicago, IL: Consortium on Chicago school research.

Torre, M., & Gwynne, J. (2009b). Changing schools: A look at student mobility trends in Chicago Public Schools since 1995. *Research report of consortium on Chicago School research at the University of Chicago Urban Education Institute.* Chicago, IL: Consortium on Chicago School Research.

Tulving, E. (1979). Relation between encoding specificity and levels of processing. In L.S. Cermak, F.I.M. Craik (Eds.), *Levels of Processing in Human Memory*. Hillsdale, N.J.: Lawrence Erlbaum Associates Inc.

Tulving, E. (1983). *Elements of episodic memory*. Oxford: Oxford University Press.

Tulving, E. (1985a). How many memory systems are there? *American Psychologist*, *40*, 385-398.

Tulving, E. (1985b). Memory and consciousness. *Canadian Psychologist, 26*, 385-398.

Tulving, E., & Schacter, D.L. (1990). Priming and human memory systems. *Science*, *247*, 301-306.

Tversky, A., & Kahneman , D. (1971). Belief in the law of small numbers. *Psychological Bulletin, 76*, 105-110.

Tversky, A., & Kahneman, D. (1973) Availability: A heuristic for judging frequency and probability. *Cognitive Psychology, 5*, 207-232.

Tversky, A., & Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases *Science, 185,* 1124-1131.

Tversky, A. (1977). Features of similarity. *Psychological Review, 84* (4), 327-352.

Wallsten, T. S. (1983). The theoretical status of judgmental heuristics. In R. W. Scholz (Ed.), *Decision making under uncertainty (*pp. 21-39). Amsterdam: Elsevier.

Waugh, N.C., & Norman, D. (1965). Primary memory. *Psychological Review*, *72,* 89-104.

Webb, A. (2002). *Statistical pattern recognition* (2nd ed.). The Atrium, Southern Gate, Chichester, West Sussex PO19 8SQ, England: John Wiley & Sons Ltd.

Wexley, K.N. (1984). Personnel training. *Annual Review of Psychology, 35,* 519-51.

Winker R.L., & Murphy, A.H. (1968). Good probability assessors. *Journal Applied Meterology, 7*, 751-758.

Wittgenstein, L. (1953). *Philosophical investigations*. (G.E.M. Amnscombe, Trans.). Oxford, UK: Blackwell publishing Ltd. (Original work published 1953).

Wyse, A.E. (2009). *A comprehensive Item Response Theory framework for evaluating standard setting*. (Unpublished doctoral dissertation). Michigan State University, East Lansing, MI.

Wyse, A.E. (in press). *An evaluation of an Angoff standard setting containing Reckase charts with comparisons between fixed effects models and correlations. Unpublished Manuscript.*

Yang, W. L. (2000, April). *Analysis of the item ratings for ensuring the procedural validity of the 1998 NAEP achievement-levels setting*. Paper presented at the Annual Meeting of the American Educational Research Association, New Orleans, LA.

Yates, J.F. (1982). External correspondence: decomposition of the mean probability score. *Organizational Behavior Human Perform, 30,* 132-156.

Yates, J.F. (1990). *Judgment and decision making.* Englewood Cliffs, NJ: Prentice-Hall.

Zieky, M. J. (2001). So much has changed: How the setting of cut scores has evolved since the 1980s. In C. J. Cizek (Ed.), *Setting performance standards: Concepts, methods, and perspectives* (pp. 19–51).Mahwah, NJ: Lawrence Erlbaum.