

Copyright
by
Winona M. Burt
2004

The Dissertation Committee for Winona Madelain Burt Certifies that this is the approved version of the following dissertation:

Connotations of Performance Level Categories Used in High Stakes Testing

Committee:

Laura M. Stapleton, Supervisor

S. Tasha Beretvas

Barbara G. Dodd

William R. Koch

Diane L. Schallert

Jay D. Scribner

**Connotations of Performance Level Categories Used in High Stakes
Testing**

by

Winona Madelain Burt, B.A.

Dissertation

Presented to the Faculty of the Graduate School of

The University of Texas at Austin

in Partial Fulfillment

of the Requirements

for the Degree of

Doctor of Philosophy

The University of Texas at Austin

August, 2004

Dedication

I dedicate this dissertation to my foundation, my family. Without their unending support I would not have had the energy to survive the dissertation process. Most importantly, I dedicate this to my parents, Winston and Gweneth Burt, who I know without a doubt would give their last breath in order to help me succeed. I am forever grateful to you both. One of the most important lessons that I have learned from my parents is that it is ok to ask for help and that it is important to be resourceful. To my sisters, Kimberley and Michelle Burt who were always there for me, thank you so much for your resourcefulness especially when I needed extra help programming my online survey. To my uncle and his wife, Carlisle and Rosie Adolphin, thank you for being my anchor and comfort. Thank you both for being there when it mattered most. Finally, to my friends, Demetrice, Michelle, and Toya, thank you so much for understanding every time I turned down offers to hangout or attend an event. Thank you for understanding and giving me the space that I needed. The dissertation process is such an enormous learning experience and through it I have learned the value of family and friends. I have accomplished this goal through the support of my family and friends and it is with their support that I will end this chapter of my life and begin a new and exciting one. As I wished for my dissertation to read like an enticing page-turner that one never wants to end, it is finished, it is complete, and it is time to move into the next phase of my life.

Acknowledgements

My dissertation would not have been completed without the assistance of Laura Stapleton, Tasha Beretvas, Diane Schallert, Barbara Dodd, Bill Koch, Jay Scribner, and Darrell Prather. It was through the tireless effort and feedback offered by my advisor, Dr. Stapleton, that I persevered and continued to strive for excellence. Dr. Stapleton's thorough feedback, words of encouragement, and support are still engrained in me. I never doubted that she was on my side and in my corner. Iron sharpens iron, and I believe that through her guidance I will continue to sharpen my blade and become a better researcher. To Dr. Beretvas, I offer my sincerest gratitude for her thorough edits and pointed questions. I have realized how difficult it can be to accomplish everything that is expected of a professor, and I recognize and appreciate your dedication to the process. To Dr. Schallert, I am so thankful that she agreed to be on my committee. Dr. Schallert was simply a delight to work with. Her enthusiasm and excitement were contagious as she helped me uncover the importance of my topic. I am also grateful to Dr. Dodd for her support and expertise. I am especially thankful that she helped me stay focused when my topic became too broad and I was overextended. Dr. Koch, thank you for your support and your expertise on the semantic differential. Dr. Jay Scribner, I am appreciative of your willingness and desire to take part in my dissertation and I wish you well. In addition to my committee members listed above, I must acknowledge Darrell Prather, a former coworker. Without Darrell's sacrifices, my online survey would not have been possible. I especially appreciate the early morning and late night programming that was necessary in order to squeeze my project into his already overcrowded work schedule. Again, thank you all who assisted and I hope the final product reflects your tremendous contributions.

Connotations of Performance Level Categories Used in High Stakes Testing

Publication No. _____

Winona Madelain Burt, Ph.D.

The University of Texas at Austin, 2004

Supervisor: Laura M. Stapleton

Cizek (1993) defined standard setting as, “the task of deriving levels of performance on educational or professional assessments, by which decisions or classifications of persons will be made” (p. 3). Much of the research in the standard setting field focuses on a compare and contrast of implementation and standard setting techniques. Nonetheless, as evidenced by the mixture of procedures implemented across the United States, researchers have concluded that there is not one “correct” standard setting procedure (Beck, 2002). In contrast, an area yet to be examined in the field of standard setting is the comparability of the performance categories employed. Selection of performance categories is one of the first tasks in the standard setting process. This task of constructing performance categories is intended to eliminate discrepancies and to facilitate understanding by each participant, the more apparent and precise the instructions and definitions, the more consistent and valid the results. The purpose of this dissertation was to investigate this key aspect of standard setting, that is, the connotation of performance categories used in high stakes testing. For example, do the performance

categories *basic*, *proficient*, and *advanced* hold different connotations than *academic warning*, *meets standard*, and *exceeds standard*? Data collection to address this and related research questions took place in two phases.

Phase one was composed of an online survey for which 167 subjects responded. Phase two of the study was composed of thirty-minute phone interviews for which four standard setting personnel participated. Study results suggested that educators perceived significant differences in the evaluative nature of the performance categories employed. For example, the term *limited knowledge* was perceived consistently less favorable than *basic* and *apprentice*. Additionally, *proficient* was preferred over *satisfactory* and *distinguished* was preferred over *advanced*. Educators also perceived differences in the level of mastery for several of the categories. However, after the provision of definitions, while significant differences in the perceived level of mastery persisted for some performance categories, these differences were lessened. As supported by each interview, these findings suggested that while connotations may at times overshadow definitions, definitions aided in mitigating these differences.

Table of Contents

List of Tables	xi
List of Figures.....	xii
CHAPTER 1: INTRODUCTION	1
Content Standards and Performance Standards	5
Standard Setting Techniques.....	6
Performance Level Categories.....	6
Study Details.....	7
CHAPTER 2: LITERATURE REVIEW	10
No Child Left Behind Guidelines	10
Summary of What States Have Done to Fulfill NCLB Requirements	12
Standards in General	19
Review of Standard Setting Methods	20
History of the Angoff Technique.....	24
General Procedures of the modified-Angoff Method	24
History of the Bookmark Method.....	26
General Procedures of the Bookmark Method.....	28
Standard Setting Summary	29
Development of performance level labels	31
Psycholinguistics in Standard Setting.....	32
Meaning According to Osgood.....	33
Meaning According to Loftus.....	35
Meaning According to Kintsch.....	36
Meaning According to Bakhtin.....	37
Summary.....	39
The Effect of Connotation on Meaning	40
The Effect of Context on Meaning	41
Statement of Problem.....	42

CHAPTER 3: METHOD	46
Phase 1: Survey of K-12 Educators	46
Purpose.....	46
Participants.....	47
Instrumentation	50
Section 1 of the Online Survey	51
Section 2 of the Online Survey	52
Overview of the Semantic Differential Technique	52
Selection of Concepts	53
Selection of the Semantic Items.....	53
Evaluation of the Semantic Differential	56
Section Three of the Online Survey	59
Section Four of the Online Survey.....	60
Instrumentation Summary.....	60
Procedures	63
Phase 2: Interviews	64
Purpose.....	64
Participants.....	64
Procedures.....	65
Analysis.....	65
CHAPTER 4: RESULTS	70
Missing Data	72
Review of Instruments	73
Research question 1: Results regarding semantic differential	73
Research question 2: Results regarding terms Measured on the no mastery mastery continuum.....	80
Research question 3: Results regarding terms with definitions provided.....	83
Research question 4: Results regarding Interviews	85
Variation in the Construction of Performance Category Definitions	86
Variation in Training and Group Discussion	90
Common Participant Questions	92

Performance Level categories and their definitions.....	93
Interview participant Opinions about performance Categories	94
Summary	95
CHAPTER 5: SUMMARY, IMPLICATIONS, LIMITATIONS, CONCLUSION	98
Overview of Results.....	99
Integrative Summary.....	101
Comparisons across States	101
Deliberation Style	102
Connotative and Denotative Meaning	104
Implications for Practice	106
Study Limitations.....	107
Future Directions	110
Conclusion	110
Appendices.....	112
Appendix A: Online Survey.....	113
Appendix B: Pre-Notification Email	128
Appendix C: Survey Launch Email	129
Appendix D: Follow-up Email.....	130
Appendix E: Follow-up Postcard.....	131
Appendix F: Interview questions	132
Appendix G: Factor Analysis Correlation Matrix and Standard Deviations	133
References.....	134
Vita	140

List of Tables

Table 2.1 Mapping of State Performance Levels to NCLB Categories.....	14
Table 3.1 List of Regions in the Sampling Frame and Their Respective States....	47
Table 3.2 Survey Return Rate for First Group of Schools.....	49
Table 3.3 Return Rate for Second Group of Schools	49
Table 3.4 Example of Typical Performance Level Categories for States.....	50
Table 4.1 Respondent Demographic Information by Region	70
Table 4.2 Survey Respondent Backgrounds	71
Table 4.3 Average Concept Score on the Evaluative Factor of the Semantic Differential	78
Table 4.4 Semantic Differential Effect Sizes.....	79
Table 4.5 Average Scores on the Mastery Continuum	81
Table 4.6 No Mastery – Mastery Effect Sizes	82
Table 4.7 Group Ratings on the No Mastery-Mastery Continuum.....	84
Table 4.8 Mastery Continuum Group Rating Effect Sizes	84

List of Figures

Figure 3.1 Online Survey Section 1 (No Mastery – Mastery Continuum)	52
Figure 3.2 Semantic Differential Adjective Pair.....	55
Figure 3.3 Online Survey Section Three (No Mastery – Mastery Continuum) with Definitions	60
Figure 4.1 Semantic Differential Factor Structure.....	74
Figure 4.2 Mean Rating for Concepts on the Evaluative Factor.....	79
Figure 4.3 Mean Rating for No Mastery-Mastery Continuum.	82
Figure 4.4 Performance Levels by Group.....	85

CHAPTER 1: INTRODUCTION

Standard setting as defined by Cizek is “the task of deriving levels of performance on educational or professional assessments, by which decisions or classifications of persons (and corresponding inferences) will be made (Cizek, 1993a, as cited in Cizek, 1993, p. 3). Typical research in the field of standard setting covers three major categories, a compare and contrast of different standard setting techniques (e.g., Green, Trimble, & Lewis, 2003; Hertz & Chinn, 2002; Reckase, 1994), an investigation of the decision making process for judges (e.g., Giraud, Impara, & Plake, 2000), and a comparison of outcomes based on the psychometric models used (e.g., Beretvas, 2004; Beretvas & Whittaker, 2002). Within these areas of research, a common theme has emerged in the literature, that is, different standard setting methods are appropriate in practice. In other words, one size does not fit all. The standard setting technique chosen should depend on the format of the test, the desired perspective for judges, and many other factors.

While the steps involved in standard setting vary depending on the technique, regardless of the technique implemented, one of the first steps, and arguably one of the most important steps in the process, is determining the number and the names of various performance categories. Nevertheless, the task of determining performance category names, in particular performance level names for statewide educational assessments, has received little attention in the standard setting community until recently. Further investigation of standard setting research reveals the lack of literature on the potential impact of the connotation of performance level categories on the placement of cutscores.

Recognizing commonalities across a range of standard setting techniques, Michael Beck of Evaluation and Testing Associates (BETA), a standard setting expert, posed several questions at the 2003 annual meeting of the National Council on

Measurement in Education. One question was whether or not the connotations of performance categories influenced the placement of cutscores in the standard setting process. Beck hypothesized that the connotations of categories do have an impact on the placement of cutscores (personal communication, 2004). Connotative meaning, as it pertains to judges during standard setting sessions, is an aspect of standard setting that could potentially influence the process and final outcome (i.e., setting cutscores). An investigation of the connotation of performance level categories and the potential impact on the standard setting process is not only critical to standardized testing in general, but it is critical to education policy considering the increased testing mandated by the No Child Left Behind (NCLB) act of 2001.

While this question of impact is an important one and one yet to be addressed in the standard setting literature, it will not be addressed here. In order to address the question of causal impact, an experimental standard setting session would be necessary and the costs and scope associated with conducting a standard setting session (even on a small scale) are beyond the scope of the current dissertation. Alternatively, as a first step, a related question will be investigated here. Specifically, the purpose of this dissertation is to determine if some of the most commonly used performance category names in standard setting hold different connotations.

The relationship between connotative (implied) and denotative (literal or explicit) meaning is made more evident through requirements introduced by NCLB. Additionally, NCLB has increased awareness and focus on testing and standard setting. The U.S. Department of Education charged each state with defining at least three levels of student performance and to specify exactly how each of those levels aligns with the *basic*, *proficient*, and *advanced* levels outlined by the U.S. Department of Education. Furthermore, NCLB mandates annual testing of all students in grades 3 through 8 and

mandates testing once during high school with the expectation that 100 percent of students will perform at the level of *proficient*, as defined by the state, by the 2013-2014 school year. However, the U.S. Department of Education allows each state to define exactly what *proficient* means; the only stipulation is that the proficient category designates passing and that there are at least a total of three categories. State flexibility in defining and naming their *proficient* level of student performance is a key source of variability across states. As an illustration, some states chose to name their *proficient* level of student performance *good*, *intermediate*, *mastery*, or *pass* all defining the same level, but each seeming to imply different meanings. Additionally, these same terms are used to decide where to set the bar (i.e., *proficient* vs. not *proficient*) which may be the single most important factor determining how states perform under NCLB.

Evidently, allowing autonomy across states has seemed to cause more divergence than convergence to a common standard. While NCLB has deemed the *proficient* level as the target standard for all students it does not define *proficient*; defining *proficient* remains the purview of each state. Not only do the terms used to describe the level of proficiency differ, but the performance and expectations of students across states vary as well. In support of this argument, student performance on the National Assessment of Educational Progress (NAEP) is offered as an example. In 2003, Bowler compared student performance on their individual statewide tests to student performance on NAEP, which is considered to be low-stakes since no student consequences are associated with the test, and found notable differences. In particular, 27 out of 29 states that administered eighth-grade reading tests reported more *proficient* and *advanced* readers on their own high-stakes state tests than what was found on the low-stakes NAEP, Louisiana, and South Carolina being the two exceptions (Bowler, 2003). This comparison demonstrates

disparities possibly stemming from differing student expectation, differing student performance, and differing denotative meaning of *proficient* across states.

Also in support of this argument (divergence from a common standard) Michael Cohen, a former assistant U.S. Secretary of Education and president of Achieve, Inc., a nonprofit organization in Washington, D.C., also acknowledged the difficulty in defining what it means to be *proficient*. State tests vary considerably in type and quality and this variety in assessments adds to the difficult task of judging what it means to be *proficient*. Cohen further adds that, “A test may be arduous...but the proficiency cutoff score may be set low so that states can easily demonstrate ‘adequate yearly progress,’ another requirement of the federal act” (Bowler, p. 2). While standards-based assessment has found its place in every state education agency due to NCLB, the expectations of student performance still vary from community to community and unavoidably so does student achievement (Bowler, 2003).

This push for higher standards across states has led to an extension of accountability to students and the associated consequences are high. For example, many states are holding students accountable by mandating that students pass a test before they are allowed to advance to the next grade. According to Olson (2003), over half of the states now require students to demonstrate what they have learned, typically in the form of a standardized test, before they receive a diploma or move to the next grade. Students as well as schools are being held accountable and are facing consequences as states strive to meet the proficiency mark.

Prior to NCLB, states voluntarily participated in NAEP but now participation is mandatory, yet consequences tied to a state’s performance on NAEP have not been set forth by the U.S. Department of Education. However, regarding student performance on each state’s test, the Department of Education made it clear that schools and districts that

fail to make adequate yearly progress (AYP) will face penalties. The percent of students who perform at the *proficient* level on a state assessment from year-to-year is the main concern of the U.S. Department of Education, and is the factor that determines if a school, district, or state is achieving adequate yearly progress.

CONTENT STANDARDS AND PERFORMANCE STANDARDS

Inconsistencies across state content and state performance standards, as a result of state autonomy, are outlined in this next section. In the context of NCLB, content standards are simply a state's specifications describing exactly what a student should know and be able to do. The development and importance of content standards are often formed with reference to external criteria, for example, the necessary knowledge at the subsequent grade, or the required knowledge for college. Stakeholders involved in establishing content standards decide what measurable behaviors a student should be able to perform in order to demonstrate a particular level of performance in a particular subject. The content standards established by each state have different foci and are a key source of variation in student performance levels across the states. Performance standards on the other hand, as related to NCLB, operationalize content standards and specify how good is good enough, and in particular how good is good enough to be deemed *proficient*. In other words, performance standards (sometimes referred to as achievement levels) communicate how well an examinee is expected to perform on a test in relation to the content standards measured by the test (Hambleton, 2001). NCLB did not explicitly mandate that each state's content and performance standards be equivalent. Nonetheless, one might argue that, to be sure "no child is left behind," state standards should have some resemblance of sameness. Otherwise, if state conceptions of *proficient* are not comparable, the intended goals of President George W. Bush's education reform might be lost.

STANDARD SETTING TECHNIQUES

In addition to the divergence in content and performance standards, the variety of standard setting techniques implemented to establish the performance standards may add to the variation in student performance seen across states. With these inconsistencies across states, it is difficult if not impossible to compare state tests that measure varying content in different manners. However, it still remains that all states must develop performance level categories and this process deserves further investigation and possible standardization.

PERFORMANCE LEVEL CATEGORIES

Another variation across states is the process used to choose their performance categories. One state for example, began by first seeking the recommendation of a technical advisory committee (TAC) to determine the number of performance categories for their statewide test. Following the recommendation of two cut points (or three performance categories) from the TAC, a second committee, a standard setting advisory committee of about 19 members was convened. This second committee was charged with determining the category names of the three performance levels for the statewide test. The names of the three performance categories along with generic definitions for each category were defined by the standard setting advisory committee, and were then used during the actual standard setting sessions. It was from this point that the various standard-setting groups, one for each grade and subject, decided what a student's performance would look like for each performance category. For this state, the names of these performance levels took the form of *below the standard*, *met the standard*, and *commended performance*.

For NAEP and 19 other states, performance levels are described with some variant of the terms *basic*, *proficient*, and *advanced*. These states have used terms that

were popular historically as performance level categories and as a result, these categories are most closely associated with NAEP. In some states, it was clearly acknowledged that the terms associated with their statewide tests were purely archival and maintained for the sole purpose of consistency. For example, for one state, the number and names of the performance levels for their state-wide test were not changed; instead after extensive discussion concerning their state and federal policy changes, constituents in the state decided to maintain their four performance levels and their four performance level names: *does not meet the standards*, *partially meets the standards*, *meets the standards*, and *exceeds the standards*. The state's policy committee had discussed the names over a six-month period and cited that the performance level categories had previously been established with extensive participation of educators and citizens. The state's policy committee further stated that the decision in 1999 and at present is to compare performance to a standard rather than to label students. Several other states simply conformed to NCLB and NAEP. Still another state, in discussing their planned levels of performance for their statewide test stated, "The proficiency levels planned include *basic*, *proficient*, and *advanced* levels to conform to NCLB requirements" (U.S. Department of Education, 2004). The connotations of these performance level categories are an aspect of the standard setting process that will be investigated here.

STUDY DETAILS

Connotative meaning is the personal meaning individuals associate with a word. In contrast, denotative meaning references the ability of a word to denote or refer to something fairly consistently, or the referential meaning of a word. Caron (1992) defines connotation as the meaning suggested by or associated with a word or an object; it can be purely individual or common to a group (Caron, 1992). According to Murphy and Zajonc (1993), connotative meaning, or whether something is seen as positive or negative, is

processed immediately by an unconscious mental system (Murphy & Zajonc, 1993). Additionally, it is a widely held belief that emotional involvement in an issue influences one's perception of that issue. Gaskins states "...emotional involvement can bias people's interpretation of an issue or event" (1996, p.386). The connection between denotative and connotation meaning in psycholinguistic literature supports further investigation of the effect of connotation in standard setting. Moreover, whether one's emotional involvement is readily apparent or masked, its effect on a person's perception is simply too important to be ignored (Gaskins, 1996).

Judges' emotional involvement in standard setting is something that should not be ignored. The words they use to communicate, chiefly the terms used in standard setting to describe student performance, all impose a point of view (Bruner, 1986). While researchers would agree that the basic function of communication is to convey meaning, it is often assumed that the connotations of these words we use to communicate elicit a similar connotative meaning between the communicator and the receiver (Osgood, Suci, & Tannenbaum, 1957). However, psycholinguistic research suggest otherwise.

Agreement on the meaning of performance level categories throughout the standard setting process is essential; otherwise, the judges using these categories to set standards will not be working from a common ground, and this could result in biased standards. Yet, even under the guidance of a facilitator and well-written definitions of each performance level category, the judges involved in the standard setting processes may likely still each hold a personal connotation and understanding of the words used to describe student performance. The purpose of the current study is to assess the connotations of performance category names used in high-stakes testing, specifically, the connotative differences in performance level category names, as well as an exploration of

the intermingling of context and connotation in the construction of meaning during standard setting sessions.

CHAPTER 2: LITERATURE REVIEW

Standard setting procedures take on many variations, they differ with regard to the types of decisions judges are asked to make, the procedure with which item difficulty is decided, the provision of student impact data, and so forth, yet, a common thread exists among them all. Each has the overall goal of determining at what point performance is considered “good enough” for passing or attaining a certain proficiency level. Under NCLB, a greater emphasis has been placed on this decision of how good is good enough. Beyond NCLB, details of standard setting procedures are important in that for some states these procedures will ultimately define the categories from which decisions on grade promotion and high school graduation will be made. Standard setting and its procedures have taken test results beyond a simple dichotomy of pass or fail. Instead, students are categorized into one of at least three performance categories, that is, a student could be categorized into three or more levels of performance. The next section presents a brief review of NCLB and a summary to illustrate states’ fulfillment of the requirements under the law, followed by an introduction to the use of standards in general. Additionally, detailed review of the two most cited standard-setting methods, the Angoff and Bookmark methods, along with the intricacies involved in choosing performance category names are also offered. Because standard setting processes involve the use of performance categories, and standard setting participant decisions may depend on understanding various terms, the final section reviews the psycholinguistic literature on wording effects. The chapter culminates in the statement of the problem.

NO CHILD LEFT BEHIND GUIDELINES

NCLB, which reauthorizes the Elementary and Secondary Act of 1965, incorporates the principles and strategies proposed by President George W. Bush. These

principles include “increased accountability for states, school districts, and schools; greater choice for parents and students, particularly those attending low performing schools; more flexibility for states and local educational agencies (LEAs) [districts] in the use of federal education dollars; and a stronger emphasis on reading, especially for our youngest children” (USED, 2001).

Increased accountability is further described to mean “...NCLB will strengthen Title I accountability by requiring states to implement statewide accountability systems covering all public schools and students” (USED, 2001, p. 3). Further, state accountability systems are expected to be based on challenging standards in both reading and mathematics. Students in grades 3-8 are expected to be tested annually and students in high school are expected to be tested at least once while in high school. Annual statewide progress reports are to be designed assessing whether all student groups (i.e., by gender, ethnicity, limited English proficiency, disability, and poverty levels) reach *proficiency* by the 2013 – 2014 school year. School districts and schools that fail to meet their statewide proficiency goals (i.e., adequate yearly progress) are subject to improvement, corrective action, and restructuring measures aimed at getting them back on course to meet state standards. Schools that meet or exceed AYP objectives are eligible for state academic achievement awards (USED, 2001).

States were required to submit a Consolidated State Application Accountability Workbook to the U.S. Department of Education by June of 2002, affirming that the state had adopted five goals and corresponding indicators and would prepare to submit baseline data in May of 2003 (USED, 2002). Performance Goal 1 is of the most interest here and it states that: “By 2013-2014, all students will reach high standards, at a minimum attaining proficiency or better, in reading/language arts and mathematics” (USED, 2002, p.11). Performance Goal 2 through 5 relate to proficient performance for

limited English learners, all students being taught by highly qualified teachers, safe and drug-free schools, and all students graduating from high school, respectively. Performance Goal 1 establishes the high stakes for schools associated with achievement at the proficiency level on states' assessments.

Under Performance Goal 1, states were also required to have defined at a minimum three categories, determined to be equivalent by the state to, *basic*, *proficient*, and *advanced* student achievement levels in reading/language arts and mathematics. For illustration, Texas' response to this goal specifies the academic achievement standards for the Texas Assessment of Knowledge and Skills (TAKS) test as: *did not meet the standard* (basic), *met the standard* (proficient), and *commended performance* (advanced). Texas terms are not similar to the terms commonly used in NAEP and many other states. The requirement that all states and all students meet the *proficient* level of academic performance is of most interest in the current dissertation research. This requirement brings what many see as 52 (including D.C. and Puerto Rico) disjointed state educational plans together for one common goal, 100% of students rated as at least *proficient* no later than the 2013 – 2014 academic year. However, the process of defining performance of a student at the *proficient* level varies from state to state; an explanation of this variation is captured in the section that follows.

SUMMARY OF WHAT STATES HAVE DONE TO FULFILL NCLB REQUIREMENTS

Every state's accountability plan was approved by the U.S. Secretary of Education, Rod Paige, after undergoing "peer reviews" by which a panel of experts reviewed the details of each plan and in many cases requested changes. Referring again to Performance Goal 1, each state was required to define at a minimum three levels of performance: *basic*, *proficient*, and *advanced*. As each state linked its performance levels to those mandated, it was common for states to include more than three levels of

performance in their accountability plans. States rationalized the inclusion of additional performance levels as sensitivity to gains at the lower levels. The more levels associated with the performance categories, the more sensitive to changes the system will be. The performance category details and the variety of terms used by the states to describe student performance are presented in Table 2.1 (USED, 2004).

Table 2.1 Mapping of State Performance Levels to NCLB Categories (USED, 2004)

State	Level A	<i>Below Basic</i> Level B	<i>Basic</i> Level C	<i>Proficient</i> Level D	<i>Advanced</i> Level E	Level F
AL		Does Not Meet Academic Content Standards	Partially Meets Academic Content Standards; Fail ²	Meets Academic Content Standards; Pass ²	Exceeds Academic Content Standards; Advanced ²	
AK		Far Below Proficient	Below Proficient	Proficient	Advanced	
AZ		Falls Far Below the Standard	Approaches the Standard	Meets the Standard	Exceeds the Standard	
AR		Below Basic	Basic	Proficient	Advanced	
CA	Far Below Basic	Below Basic	Basic	Proficient	Advanced	
CO			Unsatisfactory	Partially Proficient, Proficient	Advanced	
CT	Below Basic	Basic	Proficient	Goal	Advanced	
DE		Well Below the Standard	Below Standard	Meets the Standard	Exceeds the Standard	Distinguished
DC		Below Basic	Basic	Proficient	Advanced	
FL		Level 1	Level 2	Levels 3 & 4	Level 5	
GA			Does Not Meet Standard; Failure ²	Meets Standard; Pass ²	Exceeds Standard; Pass Plus ²	
HI		Well Below Proficiency	Approaches Proficiency	Meets Proficiency	Exceeds Proficiency	
ID		Below Basic	Basic	Proficient	Advanced	

Table 2.1 Mapping of State Performance Levels to NCLB Categories (USED, 2004)

State	Level A	<i>Below Basic</i> Level B	<i>Basic</i> Level C	<i>Proficient</i> Level D	<i>Advanced</i> Level E	Level F
IL		Academic Warning	Below Standards	Meets Standards	Exceeds Standards	
IN			Did Not Pass	Pass	Pass +	
IA			Low	Intermediate	High	
KS		Unsatisfactory	Basic	Proficient	Advanced	Exemplary
KY		Novice	Apprentice	Proficient	Distinguished	
LA		Unsatisfactory	Approaching Basic	Basic	Advanced, Mastery	
ME		Does Not Meet the Standard	Partially Meets the Standard	Meets the Standard	Exceeds the Standard	
MD			Basic	Proficient	Advanced	
MA		Warning; Failing ²	Needs Improvement	Proficient	Advanced	
MI		Below Basic, Apprentice	Basic	Met Expectations	Exceeded Expectations	
MN		Level 1	Level 2	Level 3	Level 4	Level 5
MS		Minimal	Basic	Proficient	Advanced	
MO		Step One	Progressing	Nearing Proficient	Proficient	Advanced
MT		Novice	Nearing Proficiency	Proficient	Advanced	
NE	Unacceptable	Needs Improvement	Acceptable	Good	Very Good	Exemplary
NV		Developing, Emerging	Approaches Standard	Meets Standard	Exceeds Standard	
NH		Novice	Basic	Proficient	Advanced	

Table 2.1 Mapping of State Performance Levels to NCLB Categories (USED, 2004)

State	Level A	<i>Below Basic</i> Level B	<i>Basic</i> Level C	<i>Proficient</i> Level D	<i>Advanced</i> Level E	Level F
NJ			Partially Proficient	Proficient	Advanced Proficient	
NM		Beginning Proficiency	Nearing Proficient	Proficient	Advanced	
NY			Level 1& 2	Level 3	Level 4	
NC		Level 1	Level II	Level III	Level IV	
ND		Novice	Partially Proficient	Proficient	Advanced	
OH		Below Basic	Basic	Proficient	Advanced	
OK		Unsatisfactory	Limited Knowledge	Satisfactory	Advanced	
OR	Very Low	Low	Nearly Meets	Meets Standard	Exceeds Standard	
PA		Below Basic	Basic	Proficient	Advanced	
PR			Basic	Proficient	Advanced	
RI	Little Evidence of Achievement	Below the Standard	Nearly Achieved the Standard	Achieved the Standard	Achieved the Standard with Honors	
SC		Below Basic	Basic	Proficient	Advanced	
SD		Below Basic	Basic	Proficient	Advanced	
TN			Below Proficient	Proficient	Advanced	
TX			Did Not Meet the Standard	Met the Standard	Commended Performance	
UT		Minimal	Partial	Sufficient	Substantial	

Table 2.1 Mapping of State Performance Levels to NCLB Categories (USED, 2004)

State	Level A	<i>Below Basic</i> Level B	<i>Basic</i> Level C	<i>Proficient</i> Level D	<i>Advanced</i> Level E	Level F
VT	Little Evidence of Achievement	Below the Standard	Nearly Achieves the Standard	Achieves the Standard	Achieves the Standard with Honors	
VA			Fails/Does Not Meet the Standards	Pass/Proficient	Pass/Advanced	
WA		Below Basic	Basic	Proficient	Advanced	
WV		Novice	Partial Mastery	Mastery	Above Mastery	Distinguished
WI		Minimal	Basic	Proficient	Advanced	
WY		Novice	Basic	Proficient	Advanced	

Note: Superscripts indicate separate performance category names at the high school level.

Two states qualified some seeming discrepancies in the mapping of their performance levels to USED's.

Colorado's state accountability notebook noted that, "Colorado standards for all students remain high in comparison to most states" (USED, p. 7, 2004). Louisiana's state accountability notebook cited, "Louisiana's *basic* is somewhat more rigorous than NAEP's *basic*" (USED, 2004).

Notice in Table 2.1 that the column headings range from level A to level F in order to accommodate the eleven states that defined as many as five levels of proficiency. Also, note the frequency with which the National Assessment of Educational Progress' *basic*, *proficient*, and *advanced* levels were adopted by states, specifically 34 percent of the states use NAEP categories.

While there are many facets to adequate yearly progress (AYP), the driving force is the percent of students who are performing at the *proficient* level. Beginning at the baseline year 2003, states have set annual goals to ensure that no child is left behind. Each year, for every subgroup, individual states examine their school's progress to ensure that they are on track to meet the 2013-2014 mark. Details of each state's AYP plan can be found at <http://www.ed.gov/admins/lead/account/stateplans03/index.html>. While the U.S. Department of Education has not explicitly required equivalency across state standards, it is understood that *proficient* for students in one state should not be substantially different from *proficient* students in another. The U.S. Department of Education has made explicit the use of NAEP to check the progress reported by states. NAEP will be administered in every state in grades 3 and 8, every other year, as a validation for what states report as progress (NAGB, 2002).

As displayed in Table 2.1, varieties of terms are used by the states in their standardized testing programs to describe what the U.S. Department of Education defined as the *basic* level of student performance. The alignment reflected in Table 2.1 was gathered from each state's accountability workbook as reported to the U.S. Department of Education. Examples of these terms are, *needs improvement*, *apprentice*, *approaching the standard*, *below the standard*, and *failure*. At the advanced level of student performance, while more than half of the states used the category *advanced*, the categories still vary

widely from *very good, substantial, commended performance, distinguished, to exceeds expectations.*

The assortment of terms used to describe student performance, for example the terms *adequate, proficient, satisfactory, capable, pass, and meets standards*, are often used synonymously in education policy discussion, when in fact, it is possible that these terms are not semantically equivalent (Hambleton, 2001). It would seem important then, to determine how these performance categories are perceived by those judges who use them to categorize student performance in standard setting sessions. In order for each state to fulfill the requirements of their accountability and assessment plans, performance levels were developed if not already established. These performance levels resulted from standard-setting sessions. A general review of standard setting and a summary of two of the most often used standard-setting techniques, the modified-Angoff and Bookmark procedure, are provided in the following section.

STANDARDS IN GENERAL

A standard of any type communicates “how good is good enough.” Standards have been established throughout several aspects of our lives; we have standards for drivers’ licenses, for high school diplomas, for college degrees, for restaurant cleanliness, and more. Standards in many cases are black and white -- yes or no, pass or fail, certified or not certified -- yet in some circumstances (e.g., in academics) the need arises to establish levels or gradations of what is considered “good enough.”

Setting academic standards (whether they are performance standards or content standards) involves defining the essential aspects of what and how much (of each subject) students should know. This charge is most often brought to a cross-section of the educational community who then write the standards that directly address the *how* and what of each subject. Following the development of the standards, efforts are made to

disseminate, review, and implement them, after which plans are developed further to monitor progress towards adoption and meeting the standards. Finally, to promote buy-in, states share information with the public about the standard setting process and its definitions (Improving America's School, 1996).

NCLB has implicitly defined for states what "good enough" means; student performance at the *proficient* level is good enough. Complications arise however, as there are 52, including D.C. and Puerto Rico, different interpretations of *proficient* across the states. While the goal of NCLB is obviously to leave no child behind, how do we know that entire states are not being left behind simply due to their mapping of performance categories and their interpretation of the state's *proficient* category? While the bill necessitates that all students perform at the level of *proficient* by the year 2013-2014, it leaves room for states to decide exactly what *proficient* means.

REVIEW OF STANDARD SETTING METHODS

When discussing academic standards, two types of standards are often considered: content and performance standards. Hambleton (2001) indicates that many persons, especially policy makers, fail to distinguish correctly between content and performance standards. Content standards, also known as academic standards, specify what a student should know and be able to do. On the other hand, performance standards, sometimes referred to as achievement standards, specify how a student must perform, typically on a standardized test, to be categorized into a performance level, such as *advanced*, *satisfactory*, or *limited knowledge*. Simply put, performance standards represent the level of performance examinees are expected to demonstrate. By way of example, Oklahoma's content standards in eighth-grade mathematics, as identified by the Priority Academic Student Skills (PASS), are as follows:

1. Algebraic Reasoning – The student will graph and solve linear equations and inequalities in problem-solving situations.

1.1 Equations

- a. Model, write, and solve 2-step linear equations using a variety of methods.
- b. Graph and interpret the solution to linear equations on a number line with one variable and on a coordinate plane with two variables.
- c. Predict the effect on the graph of a linear equation when the slope changes (e.g., make predictions from graphs, identify the slope in the equation $y = mx + b$ and relate to a graph) (p.2).

In contrast, the Oklahoma (2003) performance standards for the eighth-grade mathematics are defined as:

Advanced: Students consistently demonstrate a *thorough understanding* [italics added] of the knowledge and skills expected of all students at this grade level.

Satisfactory: Students demonstrate a *general understanding* [italics added] of the mathematics knowledge, skills, and processes expected of all students at this grade level.

Limited Knowledge: Students demonstrate a *partial understanding* [italics added] of the mathematics knowledge, skills, and processes expected of all students at this grade level.

Unsatisfactory: Students do not demonstrate at least a *limited knowledge* [italics added] level of the skills expected of all students at this grade level. Students scoring at the unsatisfactory level should be given comprehensive mathematics instruction. (p.1)

Notice that the degree of understanding changes from level to level; performance at the *advanced* level is demonstrated by “thorough understanding” while performance at the *unsatisfactory* level is demonstrated by “limited knowledge.” Performance standards communicate “how well” examinees are expected to perform in relation to the content or what they are supposed to know and these standards are the primary focus for the current study.

Cizek (2001) defined standard setting as “the task of deriving levels of performance on educational or professional assessments, by which decisions or classifications of persons (and corresponding inferences) will be made” (p. 3). Standard setting, then, is a method or procedure by which content standards adopted by the community are translated into performance standards (Hambleton, 2001).

Performance standards are the result of the standard setting process, and the process itself can take on many shapes and forms. Setting performance standards is a means of translating broad visions of improvement into more specific parameters for outcomes. A standard, according to Cohen, Kane, and Crooks (1999), is an “explicit decision rule that assigns each examinee to one of several categories of performance based on his or her test score” (p. 344).

Standard setting is also viewed as a process to establish buy-in for stakeholders. For some states, standard setting accomplishes three pertinent goals. First, it communicates that all students are expected to excel academically. Second, it catalyzes communication between parents and other community members about what students should know and be able to do. Third, it involves all stakeholders of the school community in the educational improvement process (Improving America’s School, 1996). Not only is standard setting a policy mechanism but also it is seen by some as parallel to the process used in the judicial system. Many researchers compare the standard

setting process to those decisions made in the courtroom (Cizek, 1993; Hertz & Chinn, 2002), that is, to the question of where to draw the line, guilty or not guilty. Regardless of one's view of standard setting, standard setting implementation is necessary in today's standards-based reform movement.

Given the general overview of standard setting, some of the specific processes involved will now be discussed. Individuals involved in standard setting procedures typically include a facilitator and a set of judges. In an effort to promote consistency, efficiency, and understanding, facilitators follow well-outlined steps when leading standard setting sessions. The facilitator is responsible for training the judges, organizing, and leading the sessions, directing discussion, and answering questions. The judges participating in the sessions are typically teachers, administrators, and community members. An example of a procedural recommendation is to have approximately 15 to 20 judges for each content area and each test for which standards are being set (Cizek, 1993). The number of cutscores set by judges is contingent on the number of performance categories desired. For example, given four performance levels such as, *advanced*, *proficient*, *basic*, and *below basic*, three cut points would be established by the judges.

While standard setting techniques abound, Livingston and Zieky outline some general consistencies across all procedures. "All procedures include the following: judges, a definition of 'borderline' knowledge and skills, procedural training for judges, collection of judgments, and combination of the judgments to choose a passing score" (1982, p. 15). Some procedures require the judges themselves to take the test. Kane (2001) states that there are at least five procedures during standard setting that could have a direct impact on the plausibility of the standards and the cutscores: (1) definition of goals for the decision procedure, (2) selection of participants, (3) training of participants,

(4) definition of the performance standard, and (5) data collection procedures. This dissertation will focus on details related to the fourth procedure, the influence the terms chosen for categories may possibly have on the overall process. Below is a review of the two most prominent standard setting methods used across all states, the Angoff and Bookmark methods.

HISTORY OF THE ANGOFF TECHNIQUE

In 1971, Angoff made mention (it was not the focus of his document) of “a systematic procedure for deciding on the minimum raw scores for passing” (p. 514). A score of one was to be awarded for each item a “minimally acceptable person” (p. 515) was judged to be able to answer correctly. The sum of the item scores would be the cutscore. What became the very widely used Angoff method was actually described in the footnote that stated “the probability that the ‘minimally acceptable person’ would answer each item correctly” (Angoff, 1971, p.515). Angoff gave no further detail on how to implement this cutscore procedure. There was no mention of how to select or train participants, and no advice was given about whether or not to allow participants to discuss their choices and revise their judgments, or whether or not to give them answer keys to the items they were judging. Because of the lack of specificity in the original description, many modern manifestations of the method allow iteration, provision of normative data to participants, and group discussion. These variants of the Angoff method fall into the generic “modified-Angoff” method nomenclature.

GENERAL PROCEDURES OF THE MODIFIED-ANGOFF METHOD

The task of judges here is to consider the item as a whole (i.e., each item separately) and to determine the probability that the “borderline test-taker” would answer the item correctly. In other words, judges determine the p-value of an item (i.e., the ratio

of test takers who answer an item correctly over the total number of test takers) with the borderline student in mind. The borderline student is conceptualized for each performance level; therefore, this step would take place for every level of performance being defined. The general procedures followed in the modified-Angoff are outlined below.

- 1) Judges either begin by examining the performance level names provided and drafted descriptions of performance levels or review the general descriptions provided.
- 2) Judges examine and sometimes take the actual test.
- 3) Round 1 begins with judges viewing one item at a time, and giving their estimate of the p-values hypothetical borderline students have (for that particular category) in answering the item correctly.
- 4) In Round 2, item judgments for the first round are discussed with the larger group and individual judges are given the opportunity to revise their original ratings.
- 5) Round 3 begins with an effort to produce convergence of item difficulty ratings. It is at this point that norming data, that is, item p-values of actual students, may be introduced.
- 6) The next step is to calculate the test score for a borderline test-taker. To do this the sum of the p-values for each item provides each judge's estimate of the borderline test-taker's expected score for the entire test.
- 7) Lastly, to produce a final cutscore, judges' expected total scores are combined by computing the mean, median, or trimmed mean. Outliers among judges are usually handled by aggregating the data using the median or trimmed mean.

The steps outlined above include some of the variations of the Angoff method. The steps call for judges to take the test (as mentioned in step 2), the provision of normative data (mentioned in step 5), or impact data. In such cases, after reviewing the data, judges would be asked to make a second probability rating of each item, which could be either the same rating or different. The recommended cutscore then would be based on this second rating. The final cutscore would be achieved in a similar manner, by summing the item probability rating for each judge to produce a total test score and then averaging the total test scores across all teachers (Buckendahl, Smith, Impara, & Plake, 2001).

An often-cited disadvantage of the modified-Angoff method is the item-by-item difficulty judgment required of participants. Shepard, Glaser, Linn, and Bohrnstedt (1993) state that "...the judgment tasks required by the modified-Angoff process we found to be difficult and confusing...the standards set seemed highly dependent on the particular sample of judges" (p. 77). Critics have also called the procedures used in the modified-Angoff method as fundamentally flawed (Shepard et al., 1993). Shepard et al. (1993) made mention to a finding of the National Academy of Education (NAE), where the panel found a general lack of consensus in interpretation of the descriptions of achievement levels that comprise the first step of the modified-Angoff method and that the descriptions were inadequate and underutilized by the judges.

The primary advantages of the modified-Angoff technique include its historically widespread use and acceptance, including its use in the development of standards for earlier forms of NAEP (Loomis & Bourque, 2001).

HISTORY OF THE BOOKMARK METHOD

The Bookmark method developed by Lewis, Mitzel, and Green (1996) of CTB/McGraw-Hill, is an item response theory (IRT)-based procedure that was developed

to accommodate changes in the testing industry. Specifically, it was developed to accommodate multiple cutscores and multiple item types (namely, constructed- and selected-response items), simplify the judgmental task by reducing and or re-focusing the cognitive load on the judges (as it allows judges to consider all of the items together as opposed to making decisions item-by-item), and connect test content with a performance level description. (Lewis, Green, Mitzel, Baum & Patz, 1999). The procedure evolved from the IRT-modified-Angoff procedure (Lewis & Mitzel, 1995). The Bookmark procedure also allows for constructed response score points to be scaled alongside the selected response score points.

The typical materials in most standard setting procedures include an operational test booklet, student papers, and scoring guides. Materials unique to the Bookmark procedure include an ordered-item booklet and an item-map-rating form. The ordered item booklet focuses the participants' attention on one item per page with the easiest item first and the hardest item last. IRT models are used to determine these item difficulties. The item map rating form is a guide to the ordered item booklet. It lists all items in the same order as they appear in the ordered item booklet, and also lists the item's scale location, the item number in the operational test booklet, the standard or objective the item measures, a space for the judge to make notes about the item, and the cutscore judgment the panelist recommends for each round. Judgments then are made at the cutscore level and not at the item level, that is, instead of making judgments about each item, judges consider all the items together to make judgments about each cutscore. The cutscore for a given performance level, for example *basic*, is identified by a bookmark placed between two items in the ordered item booklet such that from the judge's perspective, the items prior to the bookmark represent content that all *basic* students

should know and be able to do. The scale location of the item immediately prior to the bookmark is used as the operational cutscore (Lewis et al., 1999).

GENERAL PROCEDURES OF THE BOOKMARK METHOD

The fundamental tasks required of judges in the Bookmark procedure involve analyzing items to determine what they are measuring and specifying which items students in the various performance levels should be expected to respond to successfully. Typical participants in the bookmark technique include a research scientist or psychometrician, technical staff, conference manager, participants or judges, large group leaders, content leader, and table leader. When utilizing the Bookmark method, it is recommended to involve approximately 18 participants per panel; participants for a given grade and content area are then typically divided into three small groups of six each (Lewis et al., 1999). A sketch of the procedure follows.

- 1) Judges are first provided the performance category names describing the levels for which they are to set cut points. Prior to the first round of judgments, participants study the ordered item booklets within their small groups, and discuss what each item measures and why each item is more difficult than the preceding items in the booklet.
- 2) Following this discussion, participants make an individual and independent Round 1 judgment, that is, they place bookmarks that indicate the items that reflect content they expect students in each performance level to know and be able to do.
- 3) In Round 2, each small group discusses the items for which there was not consensus according to the small group's Round 1 judgment. Following the discussion, Round 1 judgments may be modified with Round 2 judgments.
- 4) Prior to Round 3, the median cutscore is calculated for each small group.

- 5) In Round 3, the large group is presented with each small group's Round 2 judgments. The median cutscore for the large group is calculated and the estimated percent of students in each performance level based on the current large group median is presented. The large group then discusses the reasonableness of this impact data and the items for which there was not consensus among the small groups.
- 6) Following the discussion, Round 2 judgments may be modified with Round 3 judgments.
- 7) Finally, performance level categories are written by the judges based on the recommended cutscores.

As outlined above, the Bookmark procedure defines performance level categories in terms of item content. Authors of the model suggest performance categories written prior to the standard setting process are ill defined because they are based more closely on the academic standards rather than the performance standards that are established as part of the standard setting process. Cut points defined based on item content are cited as a major advantage of the Bookmark method (Lewis, et al., 1999). Other advantages of the Bookmark method are that it is a whole-task method that it is based on actual student results, and accommodates multiple-choice and constructed-response items equally well.

Disadvantages of the method relate to the accuracy of scaling student results, which is dependent on the appropriateness of the IRT model used (Beretvas, 2004). In addition, often cited as a disadvantage of this method is the lack of extensive history (Kiplinger, 1997) as a result the technique is often subject to legal challenge.

STANDARD SETTING SUMMARY

Of the many steps in the standard setting process, development of the performance category names is the first to occur. Performance category names are often

seen as general and evaluative in nature, and more communicable than the test scores alone (Kane, 2001). These terms should be able to communicate to the general population the difference between each level and should make sense across content areas. Performance standards themselves have many purposes, such as (1) motivation for teachers and students (2) exemplification of achievement expectations (3) accountability for schools, and (4) certification when standards are associated with decisions for individuals (Linn, 1994). One of the most important steps in this process of standard setting is determining where cutscores (sometimes referred to as passing scores or standards) are placed.

However, the focus of the current study is not cutscores, content standards, or even performance standards. The crux of this investigation is one that has garnered little attention from the standard setting community until now, that of the performance category labels assigned to describe different student performance levels. Mehrens was quoted as saying, “The most general conclusions that can be drawn from standard setting research is that different methods produce different standards” (1995, p. 229). Green, Trimble, and Lewis (2003) in their comparison of three standard-setting procedures in Kentucky, concluded that with the diverse tasks associated with each standard setting procedure it is not surprising that different outcomes occur (Green, Trimble & Lewis, 2003). Just as different assessments measure a similar domain using various objectives and formats, the variety of standard setting procedures utilize different judgments to determine expectations for student performance and should be expected to yield dissimilar results (Crocker & Zieky, 1995).

In addition to the diverse perspectives provided by the variety of standard setting techniques, one might also expect different standards across states, considering that each state’s test measures different goals and objectives. Nonetheless, the ability of

participants clearly to conceptualize the knowledge, skills, and abilities of students within each performance level is fundamental to any standard setting process (Lewis et al., 1999).

DEVELOPMENT OF PERFORMANCE LEVEL LABELS

Selection of performance level categories in most cases is not well documented. However, in Texas, before the standard setting panel convened to set the standards, the number of cuts, the performance level labels, and their categories were decided by a separate committee (BETA, 2002). The Texas Education Agency (TEA) determined, with advice from the National Technical Advisory Committee, that two cut points should be set for the Texas Assessment of Knowledge and Skills (TAKS) tests, resulting in three levels of student performance. In addition, preceding the standard setting meeting, a standard setting advisory panel was convened with the purpose of identifying the labels for the three categories and developing generic definitions. This six-hour session was facilitated by the contractor. The session began by presenting the panel of 13 members with a broad range of choices for labels. After selection of preferred labels, ideas believed to be key for each level of performance were discussed, and generic definitions for all grades and content areas were generated (BETA, 2002).

Labels and their generic definitions as adopted by TEA are as follows: *commended performance*, performance well above the standard; *met the standard*, performance above the standard; and *did not meet the standard*, performance below the standard. This information was then provided to each standard setting panel, and they further defined the levels in terms of concrete student behaviors for their assigned grade level and content area (BETA, 2002).

Before a standard-setting group is convened, it appears to be typical that the performance level categories have already been decided, although this process is not well

documented. It is possible that the term or terms chosen to describe a performance level could influence the cutscore recommendations made by committee members. Performance labels are used to define further the expected performance of students in each category, and these classifications may play a significant role in the development of standards. Participants' attitudes and connotations associate with performance labels are important and should be investigated. In an effort to broaden understanding of the influence of words on judgment, a closer look at the psycholinguistic research on wording effects will be presented in the next section. Standard setting discourse organizes and gives structure to the manner in which student performance is to be talked about (Kress, 1989). The relative importance of this discourse, this conversation between judges in the standard setting sessions, and the conversations of those determining the performance level labels, is established in the subsequent section.

PSYCHOLINGUISTICS IN STANDARD SETTING

What we say and how we say it matters. Researchers have argued that the words we use to communicate and how they are understood by the “comprehender” is central to any investigation of meaning. Several research traditions and theoretical frameworks could be used to inform attempts to establish the importance of meaning and its related elements in the context of setting performance standards. From among these different theoretical possibilities, four have been chosen that each contribute a different lens through which to view the meaning-making process. The intent here is not to give a full explanation and review of each theory, but to present a summary of each theorist's ideas on meaning and how these ideas might contribute to the study of performance descriptors in standard setting. The four theorists chosen represent a broad range of perspectives, beginning with Osgood who in the late 50s and early 60s presented his model of meaning. Given the time period of his work, it could be argued that his theory was still

very much influenced by the predominance of the behavior theory approach to explanations of human functioning. The second theorist reviewed is Loftus who in the mid 70s reported findings from a psycholinguistic perspective showing how particular choices of words influenced the meaning individuals created as they interacted with their world. The third theorist, Kintsch, whose work was most influential in the 80s and 90s, offers the perspective that understanding involves the construction of a “situational-model” of the task or event. The fourth theorist, Bakhtin, whose original work predates that of Osgood, but who was not widely introduced to U.S. academic circles until the late 70s and 80s, offers the concept of dialogicality, or that utterances are inherently related to other utterances, and are understood by their juxtaposition with other utterances.

Most important here is the value of presenting a range of theories. Osgood (1957) clearly illustrates the need to be both selective and broad in coverage of theoretical frameworks as he stated “there are at least as many meanings of ‘meaning’ as there are disciplines which deal with language” (p.2). The choice to include these four theorists’ views of “meaning” does not imply that other meanings of “meaning” are incorrect, rather that the selection was predicated on incorporating in a wide range of theories from past to present. Following this review, an integration of theories will facilitate discussion on the importance of connotation, context, and meaning as each interacts in the context of performance level descriptors in standard setting.

Meaning According to Osgood

Charles E. Osgood, an American psychologist and communication scholar, made significant contributions in the social and behavioral sciences from the 1950s through the 1980s. Osgood is most renowned for *The Measurement of Meaning* (1957), his work with Suci and Tannenbaum in which they took an atypical approach to defining meaning. The “philosophical tradition” as recognized by Osgood and his colleagues essentially states

that meanings are infinitely variable. A researcher embracing this tradition would not readily submit meaning to measurement, because of its instability. Instead, Osgood defined meaning as a relational concept. It is because the words we use to communicate carry with them particular meanings, particular associations, and are used consistently in particular situation they reliably produce certain responses from ourselves and others. Consistency in occurrence then facilitates predictable associations with other words. In short, meanings people attribute to signs (or words) are fairly constant (i.e., at the person level), and lend themselves to measurement (Osgood, Suci, & Tannenbaum, 1957).

In their theoretical framework, Osgood et al. identified meaning as a representational mediation process, and specified objective stimulus and response conditions under which meaning is constructed. This framework was ultimately depicted through their development of the semantic differential in which a concept (stimulus), for example *feminist*, is rated by several adjective-pair items (responses) representative of the concept's meaning. While Osgood's theory of meaning is quite different from those proposed by Bakhtin, Loftus, and Kintsch, he emphasized that his theory was not meant to discount other theories of meaning (Osgood et al., 1957). Osgood further reported that he and his colleagues agreed that one of the most important factors in social activity is meaning and changes in meaning; therefore, how a person behaves in a situation depends upon what that situation means or signifies to him or her (Osgood et al., 1957).

The meaning of "meaning" for which Osgood et al., has established the semantic differential is a psychological one, and is described as a process, "That process or state in the behavior of a sign-using organism which is assumed to be a necessary consequence of the reception of sign-stimuli and a necessary antecedent for the production of sign-response" (p.9). Simply put the behavior of a person which is assumed to be necessary in order to communicate is also a necessary precursor for the production of responses.

Within the general framework of learning theory, Osgood et al. identified this cognitive state, meaning, with a representational mediation process and have tried to specify the objective stimulus and response conditions under which such a process develops.

The connotation of meaning is where Osgood (1957) found great interest. His interest in this dimension resulted in development of the semantic differential scale. The core of the semantic differential is developing a set of “polar” adjectives used to describe a concept in order to plot the differences between individuals’ connotations for words, and this same theory will be applied here. Osgood et al. using an eclectic mixture of stimuli developed and tested the theory behind the semantic differential. Chapter 3 provides more detail on Osgood’s semantic differential as a measure of connotative meaning.

Meaning According to Loftus

For over 30 years, Elizabeth F. Loftus has contributed to an understanding of human memory, most notably in her work in the field of eyewitness memory. Loftus’ work in human memory sheds light on the fluidity of what we know and what we think we know; in her research, Loftus established the importance of how questions are framed. As an example, Loftus and Palmer (1974) asked participants to estimate the speed of cars in a movie clip they had watched: “About how fast were the cars going when they *smashed* into each other?” (p. 586). Loftus and Palmer found that the verb *smashed* elicited higher estimates of speed than questions that used alternate verbs such as *collided*, *bumped*, *contacted*, or *hit*. Explanations of the higher estimates of speed that are offered by Fillmore (1971) involve specification of differential rates of movement, or that the terms used communicate differential rates of speed to the respondent. Loftus and Palmer (1974) ultimately concluded that changing a single word in a question can

markedly and systematically affect a witness' answer to a question. The authors further explained the results of their study by proposing that two kinds of information go into one's memory for some complex occurrence, that of information gathered during the original event and external information supplied after the event. Together, this information (gathered during the perception of an original event and the information supplied after the event) can in fact cause a shift in the memory representation of the incident to be more aligned to the representation suggested by the subsequent information (e.g., *smashed*). In short, questions asked subsequent to an event can cause a reconstruction in one's memory of that event.

Additional work by Loftus (1973) included the study of what was coined as the spreading-activation model of memory. The activation of an instance, or a single term (for example, *car*), simultaneously activates parallel or similar instances, such as *vehicle*. In applying this framework to eyewitness testimony, a question asked of a witness activates parallel instances. The question itself undergoes this same spreading activation, and depending on how the questions are asked, might cause the individual to adjust what he or she recalls aligning more closely to the context of the posed questions. In essence, the person might adjust his or her recollection of a situation so that it more closely relates to how the question was framed. The data of the Loftus and Palmer (1974) study supported the notion that more than one "memory location" can be simultaneously activated by a single term.

Meaning According to Kintsch

In defining discourse comprehension, Perrig and Kintsch (1985) considered three levels: a surface (text base) or verbatim representation, a propositional representation, and a situational or mental model. Comprehension at each of these three levels is associated with differential behavior. Understanding at the surface level is demonstrated

through recall; construction of adequate propositional representations is confirmed through recognition; and knowledge at the situational-model level is demonstrated in one's ability to make inferences. Perrig and Kintsch (1985) contended that according to their model of understanding, "...comprehending a text often involves the construction of a model of the situation described by the text" (p. 503). Achieving understanding at the situation-model level then is demonstrated when a reader integrates the information derived from a text with his or her prior knowledge (Kintsch, 1994).

Prior knowledge both facilitates and limits what can be acquired or understood by a learner. According to Kintsch, learning is a process that requires the active construction of a situation model and the integration of text information with the reader's prior knowledge (1994). Kintsch concluded that content overlap between text and knowledge appears to be a necessary condition for learning from text. Texts that are optimal for learning should overlap in content sufficiently, but not totally, with what readers already know. One's situational model may be sketchy or elaborate, right or wrong, but something beyond the text itself must be there in order to obtain a deeper understanding.

Meaning According to Bakhtin

Mikhail Bakhtin (1895-1975), a Russian linguist, described language (more specifically an utterance) not only as a means of self-expression, but fundamentally a phenomenon that is socioculturally situated. According to Bakhtin, utterances, "the real unit of speech communication" (1986, p.71), are inherently related to other utterances. In other words, what Bakhtin terms a "live utterance" or live speech (as opposed to the words and sentences used by linguists to support their theoretical analyses) is inherently responsive. The act of understanding utterances is filled with responses and essentially the listener becomes the speaker as understanding is sought (Bakhtin, 1986). Communication or utterances then represent a dynamic morphing of ideas between one's

self, past social interactions, the current social context, and the actual utterance itself. In our efforts to convey meaning, through text and spoken words, our dialogue directly reflects what we have already heard, read, written, and our anticipated responses. In short, our thoughts and expressions are not neutral; they are shaped by our experiences and expectations. In essence, our experiences give way to the construction of meaning. According to Bakhtin, meaning is not embedded in text (words) themselves but is constructed between people. By simply existing, we are in constant dialogue with others on a journey towards meaning (Hoel, 1997).

Meaning is negotiated through context, culture, and daily interactions, each continuously acting and reacting within the cycle. The cycle of meaning is especially apparent in group dialogue in which one's response is directly related to what others have said, read, or referenced based on their own experiences. This cycle of meaning also applies to written text which is shaped by past experiences that penetrate our minds consciously and unconsciously. It is for this reason that Hoel, in discussing Bakhtin's theory, described readers as "co-creators" of text. Readers interpret from their experiences, their purposes for reading the text, and their knowledge and associations. Hence, a text is never the same for different readers (Hoel, 1997). In Bakhtin's view, the construction of meaning is responsive and fluid. A message is not simply transmitted to the receiver; instead there is a constant interaction between the two, in effect reciprocity of ideas. "Truth is not born nor is it to be found inside the head of an individual person; it is born between people collectively searching for truth, in the process of their dialogic interactions" (Bakhtin, 1984, p.110). The construction of meaning is not a product but a continuous process. Words carry with them the places they have gone "in other people's mouths, in other people's contexts, serving other people's intentions: it is from there that one must take the word, and make it one's own" (Bakhtin, 1984, pp. 293-294).

Summary

Each theory outlined above provides a common framework for understanding the impact of connotation, context, and discourse with regards to performance level categories as they are used in standard setting. The construction of meaning, as it relates to judges' understanding of what a *borderline* or *proficient* student should know and be able to do, represents a constant dialogue that takes place throughout a standard setting session. The dialogue is not only between the facilitator and judges, but the backgrounds, the social context, reactions, expectations, and the words themselves are all pieces of the greater picture that interlock to develop meaning. When a facilitator describes to a judge the definition of a *proficient* or *basic* student, the utterances of the facilitator are bound by his or her sociocultural history, context, and the expected interaction and responses from the judges, who then in an effort to understand, must either "agree or disagree with it, augment it, apply it, and prepare for its execution" (Bakhtin, 1986, p.68), until essentially the listener becomes the speaker. Similarly, in the context of standard setting, judges communicate with the facilitator in an effort to understand their tasks and in particular to understand what each performance descriptor essentially means.

Providing judges with a definition of a performance descriptor does not guarantee that judges are all operating from the same level of meaning. As Bakhtin stated, the words themselves do not hold meaning; it is the dialogue that exists between people and their situation that establishes what one accepts as truth (1986). This search for truth between individuals reiterates the importance of the dialogue that takes place during the standard setting session as well as the dialogue that takes place when deciding on the performance level descriptors. Kintsch's theory asserts that in order for a reader to understand a text, he or she must have some prior knowledge to draw from and each theorist would support that this prior knowledge one draws from is not neutral. Not only

is the prior knowledge biased, but as judges interact with one another in the process of setting standards, meaning for one judge is manipulated, redirected, and reshaped by the input and reactions of everyone around.

Loftus realized and demonstrated the effect of word choice on memory. If the verb *smashed* produces a different recollection of an event when compared to the term *hit*, so might the term *failure* when compared to the term *novice* in describing student test performance. Beyond that, it is clear that the interaction during a standard setting session is biased to the experiences of the judges in the room, and it should also be expected that as the dialogue continues, as judges form for themselves a situational-model, or create for themselves an image of what a *basic* student should know and be able to do, this judgment is not solely based on the definition at hand. In summation, each of the theories outlined above support the idea that connotation and context each play an important role in the development of meaning, and therefore suggest that these variables must be taken into consideration to improve the process of selecting performance descriptors in standard setting.

The Effect of Connotation on Meaning

Linguists define connotation as a personal aspect of meaning that involves the emotional associations that a lexeme, the smallest meaningful unit of language, bring to mind. Connotation then is the affective meaning suggested by or associated with a word or an object and it can be purely individual or common to a group of individuals, however large or small (Caron, 1992).

The connotations of a word can be derived from background knowledge that the word invokes (Taylor, 2002). For example, Berryman-Fink and Verderber (1985) investigated the attributions and evaluative connotations associated with the term *feminist*. Labels, such as *feminist*, evolve as a convenient device for identifying and

categorizing the pro and con factions of a movement. While many labels, including “the women’s movement,” “equal rights for women,” and “women’s liberation,” all identify the movement for women’s political, economic, and social rights, not all individuals choose to associate themselves with the term *feminist*. For example, many individuals remark, “I’m not a feminist, but...” before aligning themselves with a certain position of the movement. Such association or disassociation with the label *feminist* seems to reflect the word’s varying connotative meaning. Because the term *feminist* lacks precision, individuals frequently must clarify what they mean when using the label. The choice of such labeling terminology is not a trivial matter (Taylor, 2002).

Given a psycholinguistic approach to the standard setting processes, judges involved in setting standards for high-stakes test may each hold a personal connotation of the given performance level categories. Not only might there be a preexisting connotation of the terms, but there may also be an assumption of meaning or intention. Kintsch (1978) asserts that knowledge (more specifically preexisting knowledge) makes the understanding processes “smart”; it keeps one’s thoughts on the right track and keeps us from exploring blind alleys. In particular, people are able to understand new thoughts and concepts because they know or have expectations of what is going to come next and draw on past experiences. Understanding then is expectation-based, and one’s understanding is possibly affected by both the context and the connotations of the situation. Connotations of a word can adjust one’s perception, meaning, or their situational-model of an event as proposed by Kintsch.

The Effect of Context on Meaning

The effect of context on meaning is illustrated in two classic papers from Asch in 1946 and 1948. Asch (1946) presented one group of subjects with a description of a person as being kind, wise, honest, *calm*, and *strong*. Another group was told that the

individual was cruel, shrewd, unscrupulous, *calm*, and *strong*. Both groups were then asked to write synonyms for *calm* and *strong*. The subjects given the first description took *calm* to mean peaceful, gentle, and tolerant, while subjects given the second description interpreted *calm* to mean cold, calculating, and conscienceless. Likewise, subjects given the first description interpreted *strong* to mean just, forceful, and courageous, while subjects given the second description understood it as meaning ruthless, overbearing, and overpowering (Jacobson, 1979). Loftus also supports the idea that context, particularly in the choice of words, has a significant impact on what conclusions will be drawn. It is important then to understand the context in which performance level descriptors are established and utilized.

Connotation and context are involved in the development of meaning. Burke (1965) established that “the names we give things, events, and people determine our behavior towards them” (p. xiv). Burke continued by suggesting that “words are not merely ‘signs’; they are names whose ‘attachment’ to events, objects, persons, institutions, status groups, and classes collectively soon tend to determine what we do in regard to the bearer of the name” (p. xv). The bearer of the name in the context of standard setting refers to students who represent each level of performance and the connotations of these terms and how they are interpreted in the context of standard setting is an area in need of further investigation.

STATEMENT OF PROBLEM

The choice of performance category names utilized by judges during standard setting sessions may not be a trivial matter, and some of these terms are clearly more suggestive than others. It has been argued that our choice of words is a significant and important endeavor that reveals attitudes, shapes perceptions, constitutes reality, and determines actions (Burke, 1965), especially as these terms are utilized by judges

throughout the standard setting process. While the three labels, *mastery*, *satisfactory*, and *intermediate* are all descriptions of *proficient* student performance; they do not necessarily hold the same connotations. As a result, judges may have differential attitudes toward them. Jacobson (1979) provides an example using the words “steadfast” and “stubborn.” Both words refer to not changing one’s position, but the former is perceived to involve an element of strength and is seen as a positive quality, while the latter is seen as being unreasonably unyielding and is considered a negative quality. Similarly, *adventurous* and *foolhardy* both denote risk taking, but the former is seen as being positive because of its association with glamour, while the latter is viewed in a negative light because it implies imprudence and recklessness. Clearly then, concepts, and here performance category names that are denotatively similar can be connotatively quite different (Jacobson, 1979).

The description of student behavior and the label attached to each performance level therefore, may be important considerations in establishing performance levels. Some states have chosen to avoid labels altogether and have instead numbered their performance levels – *level I*, *level II*, *level III*, and so on – to avoid any value statements and to allow more detailed descriptive statements to define what each performance level means. It is of interest then to determine the attributions and evaluative connotations associated with the terms utilized during the standard setting process. The present research was conducted to investigate just that, specifically the meanings of performance category names referenced during the standard setting process. The following four research questions were designed to assess whether there are connotative differences across seven commonly used performance level categories and explore how standard setting processes rely on these categories.

Research question 1: How do the seven selected performance level categories differ in connotative meaning? As suggested by Asch and demonstrated by Loftus, the words used to describe performance levels in standard setting might elicit different connotations, and this research question serves to assess whether judges react differently to or hold differential connotations for a selection of terms. In answering this question, the factor structure of Osgood's semantic differential scale was also assessed.

Research question 2: Is there a difference in meaning in selected performance level names when judges compare the terms on a continuum of mastery? Similar to research question 1, this question seeks to determine if the seven performance level categories hold different connotations. However, this measurement does not rely on Osgood's hypothesized underlying constructs. Instead, to address this research question judges are asked to place the performance category on a no mastery - mastery continuum to differentiate the perceived connotations for the seven categories.

Research questions 3: If definitions are provided with the performance category, are there differences in connotation of performance level categories? Also related to research question 1, this question seeks to determine if words hold different connotations; however, providing definitions for each term provides a context from which meaning or connotation can be derived. Perrig and Kintsch (1985) supported the notion that in order for people, or judges in this case, to comprehend the definitions provided in a standard setting session, each must construct a model of the situation as described by the text. At the same time, psycholinguistic theory supports that the judges each hold some preconceived ideas of what the terms should mean. The purpose of this research question was to determine whether connotation still plays a role given the same denotative (or explicit) meaning.

Research question 4: How are performance level names referenced during standard-setting sessions? Agreement on meaning across judges in a standard setting session is presumed to be critical. Standard setting personnel observations of judges' construction of meaning and dialogue that takes place during the standard setting sessions may offer insight into this key element. Interviewing standard setting personnel is a first step in exploring the role of expectations and past experiences in the construction of meaning for the standard setting tasks at hand.

CHAPTER 3: METHOD

Investigating the connotation of performance category names used in high-stakes testing was the primary focus of this dissertation. Two phases of data collection were implemented: a self-report Internet-based survey of K-12 teachers, principals, and superintendents and interviews with several participants from standard setting sessions. A review of the overarching study questions follows:

1. How do the seven selected performance level categories differ in connotative meaning?
2. Is there a difference in meaning in selected performance categories when judges compare the terms on a continuum of mastery?
3. If definitions are provided with the performance category, are there differences in connotation of performance level categories?
4. How are performance level categories referenced during standard setting sessions?

Subsequent sections of this chapter provide an overview of the purpose and describe the instruments, participants, procedures, and analyses utilized in each of the two phases of this dissertation research.

PHASE 1: SURVEY OF K-12 EDUCATORS

Purpose

This phase of the study produced data to help answer research questions one through three and focused on collecting quantitative data from educators and administrators regarding their connotative rating of various performance level categories.

Participants

Educators, in particular teachers and administrators such as principals and superintendents, typically represent over half of the participants at any given standard setting session; therefore a sample of regular instruction elementary and middle school educators was thought appropriate to represent the standard setting community. The Common Core of Data (CCD) (a comprehensive, annual, national statistical database of information concerning all public elementary and secondary schools and school districts, maintained by the U.S. Department of Education) provided current data on schools and districts across the nation and is available on a website located at <http://nces.ed.gov/ccd/>. To develop my sampling frame, states were classified into one of nine regions as listed in Table 3.1.

Table 3.1 List of Regions in the Sampling Frame and Their Respective States

Region	States
1 Middle Atlantic	NJ, NY, PA
2 New England	CT, ME, MA, NH, RI, VT
3 East North Central	IL, IN, MI, OH, WI
4 West North Central	IA, KS, MN, MO, NE, ND, SD
5 South Atlantic	DE, DC, FL, GA, MD, NC, SC, VA, WV
6 West South Central	AR, LA, OK, TX
7 Mountain Census	AZ, CO, ID, MT, NV, NM, VT, WY
8 Pacific	AK, CA, HI, OR, WA
9 East South Central	AL, KY, MS, TN

Multi-stage sampling was implemented to obtain the sample and each stage of random selection was done using SAS PROC SURVEYSELECT. First, states were stratified by region, and four school districts were randomly selected per region (two that

were categorized as rural according to CCD and two that were considered urban according to CCD). The second sampling stage involved randomly selecting two schools (elementary or middle schools only) from each district. The final sampling stage involved randomly selecting seven teachers from each school (if available). The final sample resulting from this three-stage sampling technique was composed of 504 teachers, 72 principals, and 36 superintendents for a total sample size of 612 potential responses.

The first step in data collection was to receive permission from the selected district superintendent to contact schools in the district. One week after initial emails were sent to superintendents, follow-up phone calls and emails were generated. At the completion of follow-up, 17 out of 36 districts agreed to participate in the online survey. Collection of teacher email addresses began by first collecting emails from school websites. If teacher email addresses were not available online, principals were emailed and called in order to collect a sample of teacher emails. After a few principals declined their school's participation, this process resulted in 235 teachers and principals who were contacted from 31 schools. After the first week of data collection, 70 teachers had responded to the survey. At this time, I decided to contact all of the elementary and secondary schools in the 17 participating districts instead of randomly selecting two schools. Selecting a full cluster of schools in a district was needed in order to obtain a sufficient sample size. Table 3.2 and Table 3.3 present a summary of contacts and return rates.

Table 3.2 Survey Return Rate for First Group of Schools

	# Email Sent	# Emails Bounced	# Post Cards Sent	# Post Cards Returned	Cumulative # Completed Surveys	Cumulative Return Rate
Survey Launch	235	23			70	33%
1 st Follow-up (Email)	165	19			107	49.5%
2 nd Follow-up (Postcards)			118	10	117	52%

Note: Second follow-up coincided with the initial contact of the second group of respondents.

Table 3.3 Return Rates for Second Group of Schools

	# Email Sent	# Emails Bounced	Cumulative # Completed Surveys	Cumulative Return Rate
Survey Launch	172	7	40	24%
Email Follow-up	132	5	50	30%

In total, 167 participants responded to the survey representing an overall response rate of 43%. Participants consisted of teachers (153) and administrators (14) from 48 elementary and middle schools across the United States. The 48 schools represented 18 districts, and one district was represented solely by the superintendent. He decided that the survey was too involved for his teachers, but that he would like to respond. At the completion of data collection, eight of the nine regions were represented by at least one school, Region 3 being the only region not represented.

Instrumentation

Given the interest, here in the categories used in standard setting, a list of 13 performance level categories reported to the U.S. Department of Education was originally chosen for comparison.

Table 3.4 Example of Typical Performance Level Categories for States

Performance Level Categories			
State	“Basic”	“Proficient”	“Advanced”
Arkansas*	Basic	Proficient	Advanced
Kentucky*	Apprentice	Proficient	Distinguished
Nebraska	Acceptable	Good	Very Good
Oklahoma*	Limited Knowledge	Satisfactory	Advanced
Utah	Partial	Sufficient	Substantial

Note: Asterisks denote the final states’ terms that were used.

The 13 terms presented in Table 3.4 were chosen based on three criteria. The terms representing the *proficient* level of student performance were chosen based on frequency of use. The terms representing the *basic* and *advanced* categories were selected to represent some of the diversity in terms across states (See Table 2.1 for the full table) as reported to the U.S. Department of Education. Frequency of usage was determined by sorting the 50 state performance levels on the *proficient* category. Over half the states use the term *proficient* for their middle performance level, and many states used terms similar to *good*, *satisfactory*, and *sufficient*. Once the typical terms used for the *proficient* level were identified, I investigated the *basic* and *advanced* categories and chose terms that

would represent the variety across states. Terms that were similar were grouped together and then terms were chosen that seemed connotatively different. For example, *limited knowledge* was chosen over *below basic*, since *below basic* seems so close to *basic*. The final selection criterion was that intact sets of state levels would be used. After pre-testing the on-line instrument with the original 13 concepts shown in Table 3.4 it was decided that, the survey was too lengthy and respondents were less likely to complete the survey. In an effort to reduce the burden on respondents, seven of the 13 terms were selected; the states representing the seven final terms is denoted with an asterisk in Table 3.4.

Section 1 of the Online Survey

The first section of the online instrument was composed of a no mastery—mastery seven-point continuum. Participants were asked to select a radio button along the mastery continuum, indicating the level of mastery they assumed each category indicated for each of the seven performance categories. While this section did not provide definitions for each of the seven performance categories, the third section of the online survey did and is described in detail in a subsequent section. The instrument for section 1 was used to determine the perceived level of mastery across the seven performance level categories. An example of one of the seven performance level categories that was rated is provided below in Figure 3.1. The instructions and the full instrument are provided in Appendix A.

Figure 3.1 Online Survey Section 1 (No Mastery – Mastery Continuum)



Section 2 of the Online Survey

Overview of the Semantic Differential Technique

The semantic differential was the medium used to measure connotative meaning of the seven selected performance level categories. The semantic differential is similar to a Likert-type scale, except that the semantic differential is typically separated by a seven point continuum and is anchored at each end with what is termed here as an adjective-pair item (e.g., good/bad). Charles Osgood published the first formulation of the semantic differential method in the *Psychological Bulletin* in 1952, and the first topical citation of the method occurred in 1959. The use of the semantic differential has continued over the past 45 years. While Osgood developed the method in the context of advertising and mass communications settings, its use and association has penetrated the realms of social psychology, clinical psychology, psychometrics, language, education, physiological psychology, applied psychology, and more (Finstuein, 1977). The semantic differential's connection with the domain of education includes areas such as academic achievement, special education, speech, intelligence, instructor-student perceptions, teacher concepts, developmental measures of learning, television instruction, human relations, teacher

training, and evaluation (Finstuein, 1977). The semantic differential has provided insight into the connotative meanings of concepts in the field of education with a reportedly high degree of reliability (Finstuein, 1977).

Selection of Concepts

The semantic differential is not a rigid measure, but a technique that is adaptable to a researcher's field of interest. Osgood et al. (1957) described the technique as a highly generalizable technique of measurement that must be adapted to the requirements of each research problem to which it is applied. The semantic differential does not come with a set of standard concepts and standard adjective-pair items; rather, the concepts and adjective-pairs used in a particular study depend upon the purpose of the research. The type of concept judged against a semantic differential is practically infinite; the type selected depends mainly upon the interests of the researcher.

Selection of the Semantic Items

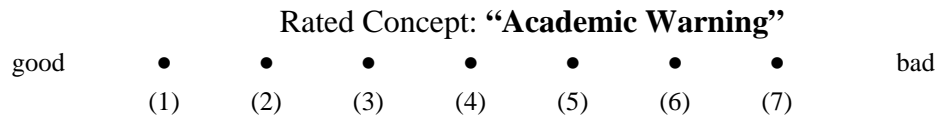
Osgood's semantic differential is composed of both concepts and adjective-pair items which are used to rate each concept. There are typically nine adjective-pair items associated with a concept, each of the adjective-pair items representing one of three major connotative dimensions proposed by Osgood, (i.e., evaluation, potency, and activity). The process of choosing adjective-pair items is more structured than that of the selection of concepts. In selecting items, small samples of closely related bipolar adjective-pair items are chosen to represent each dimension of semantic space: evaluation, potency, and activity.

Osgood (1957) wrote: "In every instance in which a widely varied sample of concepts has been used, or the concept variable eliminated as in force-choice among the adjective-pairs the same three factors have emerged in roughly the same order of

magnitude” (p. 72 – 73). First, the evaluative factor appears as the most dominant factor, followed by the potency and the activity factors in that order. In addition, it was consistently found that the evaluative factor accounted for 50 to 75 percent of the overall variance where the potency and activity factors each typically accounted for around 25 to 33 percent or half of the variance accounted for by the evaluative factor. The evaluation factor is concerned with the attitudes we attribute to something, that is, “Is it good, or is it bad?” The second factor, potency, is concerned with power and things associated with power, such as, size, weight, and toughness. The third factor, activity, is concerned with quickness, excitement, warmth, and agitation. The evaluation, potency, and activity factors of semantic space were shown to appear consistently when adjective-pair items were used to judge a concept regardless of the concept and regardless of the items (Osgood, 1957).

Adjective-pair items typically used to indicate the evaluative factor are good-bad, pleasant-unpleasant, and valuable-worthless; adjective-pairs typically used to indicate the potency factor are strong-weak, large-small, and rugged-delicate. The third factor, activity, is usually indicated by adjective pairs such as fast-slow, sharp-dull, and hot-cold (Kerlinger, 1986). Data resulting from the semantic differential is often recorded using a continuum ranging from 1 to 7, with the left side of the continuum representing positive meaning, the right side of the continuum representing a more negative meaning, and the middle of the continuum (i.e., 4) representing a neutral position. A typical adjective-pair item is shown below in Figure 3.2.

Figure 3.2 Semantic Differential Adjective Pair



In the example item shown in Figure 3.2, Osgood et al. (1957) suggested that values of 1, 2, 3, and 4 are interpreted as extremely good, quite good, slightly good, and neutral respectively, (anchor points are not visible to the respondent). Also, as suggested by Osgood et al. (1957) values 5, 6, and 7 represent slightly bad, quite bad, and extremely bad, in that order. In general then, the meaning of a concept to an individual is defined operationally by averaging the three adjective-pair items for each of the three factors, representing three scale scores for each concept. For example, if the performance category *academic warning* were the concept to be rated by a survey respondent, average scores for each of the three factors might result in 6.33 for the evaluation factor, 4.0 for the potency factor, and 7.0 for the activity factor. These scale scores would be interpreted to mean that this survey respondent perceived the term *academic warning* as quite bad, indifferently potent, and extremely passive. The meaning of a concept for a group (e.g., teachers) is determined by averaging the scores on each of the three scales for a factor which yields three averaged scale scores (i.e., evaluation, potency, and activity).

Following the recommendations of Nunnally (1967) and Kerlinger (1986), adjective-pair items were selected for this research based on two criteria. The first criterion for selecting adjective-pair items is their dimensional composition. It is recommended to select about three adjective-pair items to represent each dimension. Each adjective-pair should be maximally loaded on one factor and minimally on all other factors. Another criterion in adjective-pair selection is its relevance to the concepts being

judged. Including irrelevant adjective-pairs in a semantic differential would yield neutral judgments, and would inevitably reduce the amount of information gained (Osgood et al., 1957). Osgood et al.'s original 50 adjective-pairs provided the pool from which adjective-pairs for this study were chosen. The 50 adjective-pairs referenced represented the most frequently used adjective-pairs during the original free association trials, and were the result of Osgood et al.'s attempt to reduce the great variety of potentially usable adjective-pairs of judgment to some limited but representative number.

For the purposes of this dissertation research, nine adjective-pair items were chosen both based on their high dimensional factor loadings from past research (Osgood et al., 1957) and their relevance to the content studied in this research. For the evaluative factor, the adjective-pairs of good-bad, pleasant-unpleasant, and valuable-worthless each had high loadings of .88, .82, and .79, respectively in the original factor analysis. For the potency factor, the adjective-pairs of weak-strong, large-small, and heavy-light all had loadings of the same value, .62. Finally, for the activity factor, adjective-pairs are composed of sharp-dull, active-passive, and fast-slow, each with loadings of .52, .59, and .70, respectively. Once responses for a respondent are collected, an overall measure of meaning for each concept is calculated by averaging the responses for each factor separately.

Evaluation of the Semantic Differential

According to Nunnally (1967), the semantic differential is, "probably the most valid measure of connotative meaning available" (p.541). While the current body of literature on the construction of meaning is much different than it was 37 years ago, the semantic differential offers the researcher a potentially objective quantitative measure of meaning. The objectivity of any instrument is determined by the reproducibility of results regardless of the researcher; as objectivity is applied here it means that "...two

investigators given the same collection of check-marks and following the rules must end up with the same meanings of concepts and patterns of conceptual structures” (Osgood, 1957, p. 125).

The reliability of concept ratings on the semantic differential has also been tested and documented. Reliability of the semantic differential was described by Osgood et al. (1957) in three ways: item reliability, variable-score reliability, and concept-meaning reliability. Item reliability refers to the consistency of adjective-pair scores, for example, on a seven-point adjective-pair item, the reliability that five will be consistently selected by a subject on a given adjective-pair (e.g., bad-good) for a given concept (e.g., apprentice). Variable-score reliability refers to the reproducibility of the aggregate scores (usually an average of the adjective-pair scores) for a factor under retest conditions. By way of example, if three adjective-pair items represent the evaluation factor, the average of these three adjective-pairs produce what is referred to as the factor score, and the reliability of this score refers to the reproducibility of this aggregate factor score (Osgood, 1957). Concept-meaning reliability refers to the reproducibility of points within the semantic space with repetition of the measurement operation. Each factor score, evaluation, potency, and activity, serve to allocate the concept to a point in three-dimensional semantic space that defines its meaning (Osgood, 1957) and the reliability of this score is what Osgood refers to as concept-meaning reliability.

Osgood provided guidelines with regard to acceptable levels of reliability. In terms of adjective-pair item reliability, changes of two units or more are expected to occur less than five percent of the time (Osgood et al., 1957). The factor-score reliability was reported to change no more than 1.0 for the evaluative factor, no more than 1.5 for the potency factor, and no more than 1.3 for the activity factor at about the five percent level (Osgood, 1957). In summary, one can expect subjects’ ratings to be accurate within

a single unit of the adjective-pair, which Osgood notes as satisfactory for all practical purposes (1957).

Criticisms of the semantic differential include the bipolarity assumption of the adjective-pairs. Opponents state that the bipolar model is not a true representation of the evaluative response behaviors of subjects when responding to the semantic differential. In Gay's (1971) dissertation research he developed a unidirectional semantic differential that was designed to determine "...whether or not the subject would reflect the assumed bipolarity when it was not built into the measuring instrument" (p. 51). Gay's results indicated that the "two ends of the evaluative scales [adjective-pair] of the standard semantic differential are neither bipolar, nor orthogonal; they are essentially unipolar, with performance on one giving positive prediction of the performance on the other" (p. 52). Gay further remarked that the separation of the semantic differential adjective-pair into single unipolar scales permits a more sensitive measure, and provides for separation of the evaluative factor from measures of response intensity. "By forcing subjects to respond on bipolar scales [adjective-pairs], we do not prove bipolarity. If subjects respond in bipolar fashion under conditions which permit – but do not force – bipolarity, this could be taken as support: however, in this study, the subjects were free to treat the stimulus words as either bipolar or non-bipolar, and treated them as non-bipolar" (p. 54).

Osgood et al. (1957) admit that they have yet to address whether or not the polar terms are true psychological opposites. However, they stated, while unidirectional adjective-pairs might serve as well as the bipolar adjective-pairs and might eliminate this problem, it would probably create another, "...if there is a 'natural' human tendency to think in terms of opposites, the so-called neutral point at one extreme of unidirectional scales [adjective-pairs] would probably tend to take on the semantic properties of oppositions" (p. 328). Considering the widespread use and documented reliability of

responses to the original semantic differential, the form initially proposed by Osgood et al. will be implemented here in order to determine if there is a connotative difference in meaning between educators and across the seven performance level categories.

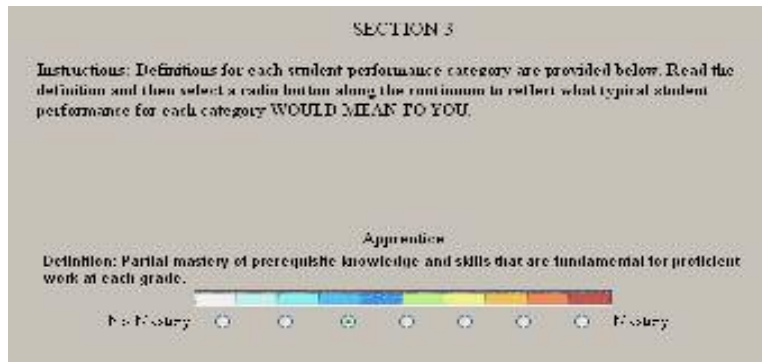
Section Three of the Online Survey

In addition to assessing the connotation of the seven performance level categories via the no-mastery/mastery continuum and the semantic differential, participants were divided into three groups to determine whether the provision of definitions alleviated any observed differences in connotation of terms. Each group was asked to rate the performance level categories for one of three states (Arkansas, Kentucky, or Oklahoma) as referenced in Table 2.1. After sampling, participants were pre-assigned to one of the three groups. The respondents in each group rated three intact state performance level categories on a no mastery—mastery seven point continuum. The definitions for each of the three levels were consistent across all three groups and were the definitions for the eighth-grade NAEP Reading Performance Levels. Section three of the online survey helped to determine whether any differences in performance levels that were found in section 1 of the survey persisted once definitions were provided.

An example of the instrument is presented below in Figure 3.3 and the full instrument is presented in Appendix A.

Figure 3.3 Online Survey Section Three (No Mastery – Mastery Continuum) with

Definitions



Section Four of the Online Survey

Section four of the online survey contained demographic questions. Data from the demographic page allowed me to determine how representative the survey respondents were of the target population. The first question determined if the survey respondent had participated in a standard setting session before. The second demographic question collected information regarding years of teaching experience. The final question collected data on the different subjects taught. All demographic information was collected in order to report sample representativeness.

Instrumentation Summary

Each participant was presented with four sections of the online instrument in the following order: the no mastery-mastery continuum, the semantic differential, the no mastery-mastery continuum with definitions, and the demographic section. Survey order was explicitly determined to ensure that performance category definitions would be presented at the end of the survey. The intent was to prevent influence of the definitions on other responses. In the first section, the survey respondents were asked to

rank each of the seven terms on a continuum. Each respondent was presented with the same seven terms in the same order. Raw data yielded a number from 1 to 7 on a continuum for each of the terms, and provided the researcher with data to answer research question 2, that is, is there a difference in meaning in selected performance level categories when judges compare the terms on a continuum?

The second section of the online survey displayed, on seven separate screen pages, each of the seven concepts for the semantic differential. It was originally planned that the concepts were to be randomly ordered for each participant as they logged onto the website. Randomly ordering the presentation of concepts would prevent consistent order effects. However, due to difficulty of working with the dynamic nature of the Internet, it was decided instead to design seven different orders in which to display each of the seven performance level categories for the semantic differential. Establishing seven predefined orders eliminated the potential browser problems when webpage order is decided dynamically. The order of the seven categories was determined so that each of the terms was presented in each position among the other words in the group.

Sequence of the nine adjective-pair items for each performance level category was consistent on each screen. The first three adjective-pair items were positioned so that positive terms were displayed on the left; this first set of three adjective-pairs represented the evaluative factor. The next set of adjective-pair items represented the potency factor and the positive terms were positioned on the right; the final three adjective-pair items represented the activity factor and the positive terms were positioned on the left. As mentioned previously each adjective-pair item was separated by a seven-point continuum. The full instrument is provided in Appendix A. Raw data produced by the semantic differential included ratings from one to seven on each of the nine adjective-

pairs for each of the seven performance level categories. In summary, each subject produced 63 points of data on the semantic differential.

The third section of the online survey was designed to elicit ratings for three sets of state terms (i.e., Arkansas, Kentucky, and Oklahoma); each state had three levels of student performance. Survey respondents were sequentially divided into the three groups (group assignments took place after the original sample was drawn). Only the terms that map onto the NCLB concepts of *basic*, *proficient*, and *advanced* levels were used. The definitions for each of the performance level categories were consistent across each of the three groups, and are the same as the policy definitions used for the eighth-grade NAEP Reading test. In this section, performance categories were provided in the same order for each participant, that is the terms representing the basic category first, the proficient category second, and the advanced category third.

The final section of the online survey, section 4, simply requested demographic information from teachers. Such items include whether or not they had participated in a standard setting session before, what subject or subjects they taught, and how many years of teaching experience they had.

Originally it was planned for the first and third sections of the online survey to ask participants to rate each of the categories on a continuum from 0 to 100 indicating the percent mastery for a student at each level instead of the 7 radio buttons. It was also planned to utilize a slider scale so that participants could slide a pointer from one end of the continuum to the next indicating a percent of mastery. Further investigation of a slider scale revealed the need for complicated FoxPro programming, and I was unable to locate a programmer who was proficient with this language under the resource constraints of this research. During survey development, I also contemplated using 33 radio buttons to represent the no master—mastery continuum, but 33 radio buttons proved overwhelming

for teachers during pre-testing. The seven-point continuum was decided on as an alternative and was found acceptable during pre-testing.

Procedures

A sampling frame of public elementary and middle schools across the U.S. were downloaded from the Common Core of Data, a database provided online by the National Center for Educational Statistics (NCES). As described earlier, multistage sampling provided the final sample of educators and administrators

Pilot testing of all instruments and communication material was initiated on Monday, February 16th and continued for five weeks through March 22nd. The online survey was piloted with four teachers and six community members. Cognitive interviews were conducted with the teachers, in which the researcher visited the school and the teachers responded to the on-line survey while verbalizing their thoughts and processes. The six community members responded to the survey on-line and provided post hoc feedback. Website functionality and data collection were tested on various monitors, browsers, and operating systems.

The complete instrument was administered online and was launched on Tuesday, March 30th, 2004 and data collection ceased on Friday, April 23rd. A dynamic link to the online instrument was distributed via email to teachers, principals, and superintendents from selected schools. A pre-notification email was sent four days in advance of launching the online survey alerting teachers to the request for their participation. The pre-notification email and the announcement email are provided in Appendices B and C. Follow-up emails were initiated one week after the launching of the survey (see Appendix D). The second follow-up came in the form of post-cards sent to the school (see Appendix E). The second group of teachers surveyed only received one follow-up via email. From launch to close, the online instrument administration took approximately

four weeks. As an incentive, teachers, principals, and superintendents were offered the chance to win one of three \$50 gift certificates in a raffle. The first teacher gift certificate was awarded on Monday, April 5th; the second was awarded on Monday, April 12th, and third was awarded on Monday April 19th.

PHASE 2: INTERVIEWS

Purpose

Assuming that performance level categories are found to be connotatively different, it would be important to know how performance level categories are referenced during standard setting sessions. If the terms are referenced more often than the definitions themselves, then there is a potential for purely connotative impact on the placement of the final cutscores. Exploratory interviews with standard setting personnel provided information with which research question 4 was addressed: How are performance level names referenced during standard-setting sessions?

Participants

Originally, I had planned to interview judges (e.g., teacher, principals, and superintendents) who had previously participated in standard setting sessions. However, sources that were able to provide names and contact information for potential participants were reluctant to do so. Instead, a group of standard setting personnel was interviewed. While changing the nature of the sample still provided information to answer the overarching research question, interviewing standard setting personnel offered a different perspective and limited the generalizability of the conclusions drawn from the interviews.

Interview participants represented three different types of standard setting personnel: facilitators, observers, and technical staff. The facilitator is typically responsible for organizing and running the session. The observer is familiar with the

standard setting process and simply monitors the standard setting session, and technical staff is responsible for running any necessary statistics and other tasks as well as observing the process. Three of the interviews provided feedback regarding standard setting for what will be referred to here as Test A, and the remaining interview provided perspectives regarding standard setting on Test B. These interviews offered insight into some of the situational models judges may create when defining their own meaning of the performance category terms they used during the various standard-setting sessions. Two of the interview participants were technical staff, one participant represented an observer, and the final participant was a facilitator. Each of these participants completed a 30 minute phone interview.

Procedures

Six possible participants were identified through conversations with psychometric experts. These possible participants were then sent an email. Four consented to participate and we scheduled times for individual phone interviews. (Interview questions are presented in Appendix F). Interview questions were first pilot tested with an employee of a state education agency. The final interview questions were refined, some were deleted, and some were added as a result of the pilot test. Interviews took place from April 9th through April 19th. All interviews were audio recorded.

Analysis

Data collected in phase one of the study were analyzed using a series of mean comparisons. Data collected in phase two of the study, the exploratory interviews, were simply summarized. Each analysis is outlined in detail in the subsequent paragraphs. In short, one repeated measures MANOVA, one repeated measures ANOVA and an independent samples ANOVA were conducted, for research questions 1, 2, and 3

respectively. Following significant findings from the MANOVA or ANOVAs, more specific comparisons, that is, comparisons by performance level were carried out.

Research Question #1: How do the seven selected performance level categories differ in connotative meaning? Using data from Phase 1, a confirmatory factor analysis (CFA) was conducted as a preliminary step. CFA was used to seek support as to whether semantic differential items are measuring the three dimensions they are reported to measure. Verification of this factor structure was essential before proceeding with further data analysis using results from the semantic differential. Provision of support for the hypothesized factor structure allowed aggregation of adjective-pair items for scale scores within each of the three dimensions.

Mplus (Muthén & Muthén, 1998) was utilized to perform confirmatory factor analysis using a maximum likelihood estimation procedure. Input data consisted of raw data from the nine adjective-pair items (three adjective-pairs per dimension). Each adjective-pair was scored on a seven-point continuum ranging from one to seven. A complication in this analysis is that each respondent rated seven concepts providing observations nested in each person. To account for this dependency, the analysis TYPE =COMPLEX was used to produce accurate standard errors and chi-square test statistics.

Results of the CFA were evaluated in terms of the chi-square statistic the comparative-fit index (CFI) and the standardized-root-mean-square residual (SRMR). As recommended by Hu and Bentler (1999), a joint criterion of values greater than or equal to 0.96 for the CFI and values less than or equal to .10 for the SRMR will be considered an indication of good fit.

Contingent on adequate data-model fit, scale scores for each selected performance category term on each of the three dimensions were calculated. However, due to the

nested nature of the data (teachers nested within schools) the amount of dependency within school was first determined by calculating the intraclass correlation. Results of the intraclass correlation aided in determining the plan of action to compensate for the clustering problem. Specifically, if it were necessary a smaller alpha level would be utilized to compensate for negatively biased standard errors.

Given the multiple factors of the independent variable (i.e., the seven performance categories), the three dependent variables (i.e., the three factors evaluation, potency, and activity) of interest, and the repeated nature of the responses (i.e., each subject rated each concept) it was decided that a repeated measures multiple analysis of variance (MANOVA) was the appropriate analysis for the data yielded by the semantic differential. Following significant results for the repeated measures MANOVA, comparisons using univariate t-tests for the three terms at the *basic* level, two terms at the *proficient* level, and two terms at the *advanced* level were used to determine which pairs of means were significantly different. The following paragraph describes the planned comparisons in detail.

First, for the three terms at the *basic* level (*basic*, *limited knowledge*, and *apprentice*) univariate t-tests were used to determine which terms were significantly difference from each other. Specifically, the scale scores for each of the three were compared to each other. The terms representing the proficient level (i.e., *proficient* and *satisfactory*) were compared to each other, and the scale scores representing the terms for the advanced level (i.e., *distinguished* and *advanced*) were compared to each other using univariate tests. Each of the post-hoc comparisons were analyzed using the most conservative measure, Scheffé, to control for the inflated experiment-wise error rate.

Research Question #2: Is there a difference in meaning in selected performance categories when judges compare the terms on a continuum of mastery? To begin, general repeated measures ANOVA addressed the question: Is there a significant difference in continuum location across the seven selected performance level categories? Following a significant finding, post hoc t-tests were utilized to determine which mean pairs were significantly different within the three levels of performance. Specifically, the mean continuum location of the terms representing the *basic* category (i.e., *apprentice*, *basic*, and *limited knowledge*) were compared to each other. Then, the terms representing the *proficient* category (i.e., *satisfactory* and *proficient*) were compared, and finally the terms representing the *advanced* category (i.e., *distinguished* and *advanced*) were compared. Following findings of significant differences for the *apprentice*, *basic*, and *limited knowledge* terms, pairwise t-test were conducted. The Scheffé test was used for each post-hoc comparison to control for experiment-wise Type I errors.

Research Question #3: If definitions are provided with the performance category, are there differences in connotation of performance level categories? Respondents were split into three separate groups and provided with three states' terms. The research question addresses differences across groups. Three ANOVA addressed the question of interest here: Is there a difference in the perceived level of mastery across the three terms selected to represent the basic category (*apprentice*, *basic*, and *limited knowledge*) when definitions are provided? Is there a difference in the perceived level of mastery across the two terms selected to represent the proficient category (*proficient* and *satisfactory*) when definitions are provided? Finally, is there a difference in the perceived level of mastery across the two terms selected to represent the advanced category when definitions are provided (i.e., *distinguished* and *advanced*)? Following significant differences for the

apprentice, basic, and limited knowledge terms, the Scheffé test was used for each post-hoc comparison.

Research Question #4: How are performance level categories referenced during standard setting sessions? Using data from Phase 2 (the exploratory interviews), comparisons of interviewee responses addressed research question 4: How are performance level names referenced during standard-setting sessions? Referencing interview questions developed through pre-testing, commonalities and differences across responses were collected. Interpretation of interviews consisted of a summary of frequent responses, noted patterns, and unique responses. A summary of the interviews helped to determine how performance level names were referenced during two different standard-setting sessions.

CHAPTER 4: RESULTS

The purpose of this study was to assess the connotation of performance level categories used in high-stakes testing. Using an Internet survey, teachers, principals, and superintendents from across the country provided their perceptions on seven of the most commonly used performance level categories. A total of 407 subjects were contacted and 167 subjects responded to the online survey for a final response rate of about 41 percent. Table 4.1 provides descriptive information regarding the demographics of the 167 survey respondents by region. Additionally, Table 4.2 provides a detailed summary of survey respondent backgrounds. In short, the majority of the teachers and administrators were from rural elementary schools, and on average they had about 16 years of teaching experience that ranged from as few as 1 year of experience to 44 years of experience.

Table 4.1 Respondent Demographic Information by Region

Region	State	Number of Districts	Number of Schools	Number of Teachers	Number of Administrators
1	NY	2	2	6	1
2	CT	1	5	13	2
2	NH	1	2	3	-
4	IA	2	9	28	3
4	NE	1	1	4	-
5	VA	1	1	5	-
6	OK	1	3	13	1
6	TX	2	8	30	2
7	ID	1	3	12	-
7	NM	1	1	1	1
8	CA	3	10	23	3
9	AL	1	3	15	1
Totals:	12	18	48	153	14

Note: Region 3 was not represented.

Table 4.2 Survey Respondent Backgrounds

School Location	Percent
Urban	46
Rural	54
School Category	
Elementary School	56
Middle School	46
Subjects Taught	
English	88
Math	77
Science	80
Social Studies	79
Other	62
Previous Standard Setting Participation	
Yes	17
No	83
Note: Please note that the “Subject Taught” categories are not mutually exclusive.	

In terms of analysis with regard to survey responders and non-responders, results of an independent measures chi-square test show that response to the survey was not dependent on teachers’ rural or urban status where $\chi^2 (1, N = 354) = 3.76, p > .05$. However, it was found that response to the survey may have been dependent on teachers’ grade level $\chi^2 (1, N = 354) = 30.01, p < .05$ where more elementary teachers tended to respond than did middle school teachers.

In addition to the online survey, interviews were conducted with four standard setting participants. Each interview participant was involved with at least two separate standard setting sessions, and their experiences in the field of standard setting ranged

from two to fifteen years. Presentation of the results will take place in two parts. Data obtained via the online survey will be presented first, followed by a summary of general themes extracted from the four individual phone interviews.

The key research questions addressed are:

1. How do the seven selected performance level categories differ in connotative meaning?
2. Is there a difference in meaning in selected performance categories when judges compare the terms on a continuum of mastery?
3. If definitions are provided with the performance category, are there differences in connotation of performance level categories?
4. How are performance level categories referenced during standard setting sessions?

MISSING DATA

The final sample size for analysis was 167; however, some data were believed to be missing at random. Using listwise deletion as a solution for missing data is not optimal; however, data were reduced by less than 10 percent when listwise deletion was invoked for a missing response. Please note that this method of handling missing data resulted in sample sizes that vary across each analysis. For instance, if only 164 people had complete data for items on the semantic differential, then all subsequent analyses were run with a sample size of 164. In contrast, if no missing data were found for the items on the mastery continuum then 167 was used as the sample size associated with these analyses.

REVIEW OF INSTRUMENTS

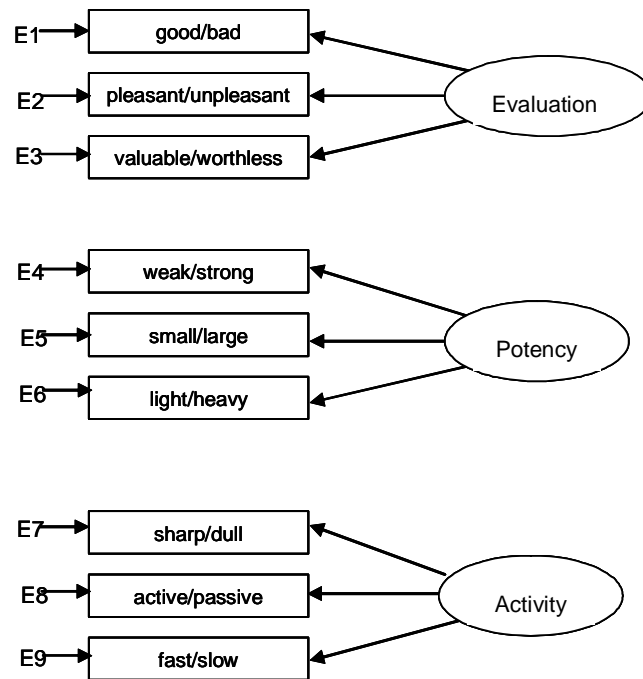
An online survey composed of four sections provided data for the first set of analyses. The full instrument is presented in Appendix A. The first section of the online survey was composed of a no mastery-mastery continuum. Here participants were presented with seven performance categories and asked to rate the categories according to their perceived level of mastery on a 7-point scale, where a 1 indicated no mastery and a 7 indicated mastery. In addition, the seven performance categories were presented in the same order to each survey respondent. The second section of the online survey was composed of the semantic differential. Typically, the semantic differential represents three dimensions of connotative meaning; evaluation, potency, and activity each of these dimensions of meaning was represented by three adjective-pair items which yielded a total of nine adjective pair items for each concept. For each adjective pair item on the semantic differential, the most positive responses received a 1 and the most negative responses received a 7. Presentation of the terms here took on seven different orders. The seven orders allowed each concept to be displayed in each position at least once. The third section of the online survey presented each participant with one of three sets of state terms coupled with definitions that were held consistent across each of the three groups. Respondents were asked to indicate the perceived level of mastery on a 7-point scale where 1 indicated no mastery and 7 indicated mastery. The fourth section of the online survey collected demographic information all of which was summarized in Table 4.2.

RESEARCH QUESTION 1: RESULTS REGARDING SEMANTIC DIFFERENTIAL

Research question 1 was an investigation of the connotative meaning across the seven performance level categories. Prior to examining this research question, a confirmatory factor analysis was undertaken using MPlus software (Muthén & Muthén, 1998). The factor structure hypothesized by Osgood et al. (1957) for the semantic

differential is presented in Figure 4.1, where ellipses represent latent factors and rectangles represent the adjective-pair variables. Arrows connecting factors with variables represent factor loadings. Input data for the factor analysis was restructured in a manner such that each respondent was represented by seven rows and nine columns. The seven rows represented a set of responses for the seven concepts, and the nine columns represent responses to the nine items for each concept. Restructuring the data in this manner resulted in a nested structure (concepts within respondent). To address this nesting, the estimation option, TYPE = COMPLEX was used in MPlus to provide accurate standard errors and chi-square statistics.

Figure 4.1 Semantic Differential Factor Structure



The good/bad, pleasant/unpleasant, valuable/worthless items were hypothesized to load on the evaluation factor. The weak/strong, small/large, light/heavy items were

hypothesized to load on the potency factor. The sharp/dull, active/passive, fast/slow items were hypothesized to load on the activity factor. All models were run under the assumption that the three factors were independent. Osgood et al. (1957) expected that a concept, in this case the performance level category, could be rated on each of nine scales with each factor being independent of the others. Osgood offered as an example, the terms hero and pacifist: “To put the matter yet another way, some of the things judged ‘good’ may also be judged ‘strong’ (e.g., hero) but other things judged equally ‘good’ may also be judged ‘weak’ (e.g., pacifist)” (1957, pg. 72).

Due to the violations of normality determined mostly by examination of univariate skewness, maximum likelihood estimation (MLM) was also implemented in Mplus; it produces robust standard errors (with regards to violations of normality) and a mean-adjusted robust chi-square test statistic. Results of the CFA indicated poor fit of the model to the data, as the chi-square value was determined to be significantly different from zero $\chi^2(27, N = 1159) = 1323, p < .05$. Additional fit indices indicated that the data did not fit the hypothesized model, with a comparative fit index (CFI) of .18 and a standardized root mean square residual (SRMR) of .31.

Consequently post hoc modification indices were inspected. Specifically, the LaGrange multiplier test indicated that fit would be moderately improved if the model was re-specified to allow the three factors to co-vary. However, considering Osgood’s theory did not support model re-specification the model was not adjusted. Instead, I was interested in inspecting model fit for the most dominant factor, the evaluative factor. One additional model was run in MPlus specifying one factor, with only three adjective items loading on it, that is, good/bad, pleasant/unpleasant, and valuable/worthless. The adjective pair item loadings on the evaluative factor were good/bad (.263), pleasant/unpleasant (.866), and valuable/worthless (.949). The quality of the just-

identified model of the evaluation factor was indicated by the variance extracted = .569 and the construct reliability .77. Generally, it is recommended that a factor account for at least 50 percent of the variance, as is the case here, and that construct reliability is at least .70. As a result, the evaluative model was used from the semantic differential responses to calculate scale scores (i.e., the average of the three adjective pair items on the evaluative factor). The correlation matrix for each of the seven concepts and their standard deviations are presented in Appendix G. Due to the reduction of data from three dependent variables to one dependent variable (i.e., the evaluative factor) the repeated measures MANOVA was reduced to a repeated measure ANOVA.

Analyses of data yielded from the semantic differential consisted of four repeated measure ANOVAs conducted using SAS. The first omnibus ANOVA tested for differences across all seven concepts. Following a significant finding across all seven concepts, comparisons were made within each of the three performance levels. The repeated factor accounted for each subject responding to all seven concepts, the independent variable was concept, and the dependent variable was the average evaluative scale score that ranged from 1 to 7 for each concept. In addition, as noted in chapter 3, given the nested structure of the data, that is, teachers within schools, intraclass correlations were inspected (as displayed in Table 4.3).

Given the results in Table 4.3, it was decided that a violation of the independent observation assumption was not an issue. The second ANOVA assumption (according to Stevens, 2002) is that of homogeneity of variance. However, according to Stevens (2002), violations of this assumption are robust with respect to Type I errors. The final assumption, that the dependent variables are normally distributed, is demonstrated via the details on the skewness and kurtosis of each dependent variable in Table 4.3.

Following significant results from the first repeated measures ANOVA, comparing means across all seven concepts, $F(6, 978) = 304.53, p < .05$, comparisons between the categories representing each of the three levels of performance were conducted (means are depicted in Figure 4.2). Results revealed a significant mean difference in perception among the terms at all three levels of performance. Specifically, educators perceived a significant difference on the evaluative scale between the terms representing the “basic” level of student performance where *apprentice* was significantly different from *limited knowledge* $t(163) = 15.14, p < .05$, *basic* was significantly different from *limited knowledge* $t(163) = 13.53, p < .05$, and *apprentice* was significantly different from *basic* $t(164) = 5.56, p < .05$. Significant difference was also found for the terms representing the “proficient level” of student performance $t(164) = 10.69, p < .05$, and between the terms representing the “advanced level” of student performance $t(164) = 7.42, p < .05$. Also, note the large effect size differences displayed in Table 4.4 for each of the terms at the basic and proficient levels of student performance, where as, the effect size at the advanced level is much smaller.

Table 4.3 Average Concept Score on the Evaluative Factor of the Semantic Differential

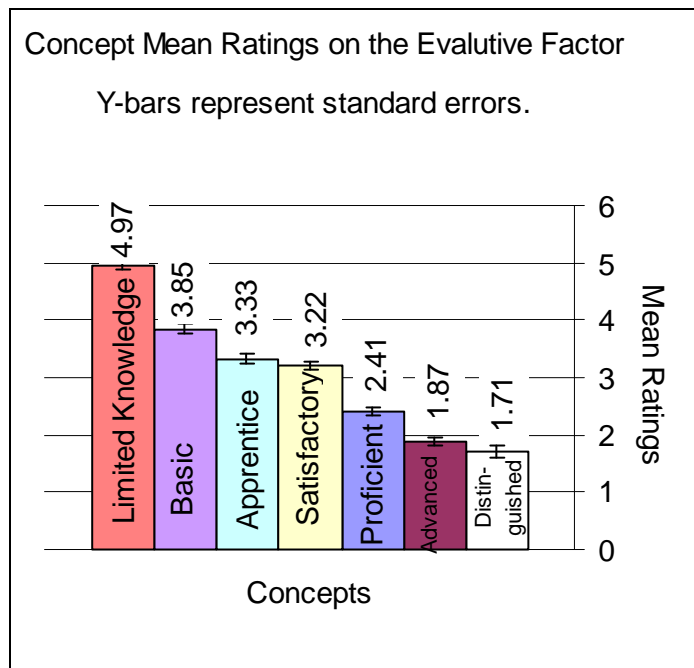
Concept	N	Mean	SD	Skew	Kurtosis	ICC
“Basic Level”						
Basic	164	3.85 ¹³	1.08	-.42	.68	-.08
Apprentice	164	3.33 ²³	1.13	-.27	-.15	.06
Limited Knowledge	164	4.97 ¹²	1.07	-.71	1.18	.01
“Proficient Level”						
Proficient	165	2.41 ⁴	.999	.04	-1.15	-.05
Satisfactory	165	3.22 ⁴	.841	-.89	.65	-.03
“Advanced Level”						
Advanced	165	1.87 ⁵	.897	1.47	3.52	-.12
Distinguished	165	1.71 ⁵	.877	1.13	.38	-.08

Note: Lower means indicate connotations that are more positive. Superscripts indicate significant pairwise comparisons at the $p < .05$ level. Intraclass correlations do not include responses at the district level (i.e., superintendents).

Table 4.4 Semantic Differential Effect Sizes

Performance Category Comparison	Cohen's d
Limited Knowledge-Apprentice	1.49
Limited Knowledge-Basic	1.04
Basic-Apprentice	0.47
Satisfactory-Proficient	0.88
Advanced-Distinguished	0.18

Figure 4.2 Mean Rating for Concepts on the Evaluative Factor



RESEARCH QUESTION 2: RESULTS REGARDING TERMS MEASURED ON THE NO MASTERY-MASTERY CONTINUUM.

Research question 2 focused on whether there was a difference in meaning in selected performance level labels when judges rated the terms on a no mastery to mastery continuum. Four repeated measure ANOVAs were conducted using SAS, that is, one overall ANOVA and one for each of the three performance levels. Prior to the analysis, intraclass correlations, skew, and kurtosis were inspected (as displayed in Table 4.5) to assess violations of observation independence and normal distribution of the dependent variables. Nonetheless, according to Stevens (2002), violations of univariate normality are robust with respect to Type I errors and ANOVAs are also robust with respect to violations of normally distributed dependent variables as long as group sizes are equal or approximately equal with the largest group differences being less than 1.5 which is also the case here.

Following significant results from the first repeated measures ANOVA comparing means of all seven performance level categories, $F(6, 984) = 301.26, p < .05$ planned comparisons between the categories representing each of the three levels of performance were conducted (means are depicted in Figure 4.3). Results revealed a significant mean difference in perception of mastery among the terms at each of two levels of performance. Specifically, educators perceived a significant mean difference between the terms representing the *basic* level of student performance where *basic* was significantly different from *limited knowledge* $t(166) = 15.55, p < .05$, *apprentice* was significantly different from *limited knowledge* $t(166) = 13.7, p < .05$, and *apprentice* was significantly different from *basic* $t(166) = 3.43, p < .05$ where *limited knowledge* had the lowest mean. In addition, with the terms representing the *proficient* level of student performance educators perceived a significant difference between *satisfactory* and

proficient $t(164) = 12.52, p < .05$. Educators did not perceive a significant difference in terms that represent the *advanced* level of student performance $t(166) = 2.32, p > .05$ that is *advanced* and *distinguished*. Note in Table 4.6 the twodarge effect size differences for *apprentice* vs. *limited knowledge* and *basic* vs. *limited knowledge*. The *proficient* vs. *satisfactory* comparison was considered small according to the criteria outlined by Cohen.

Table 4.5 Average Scores on the Mastery Continuum

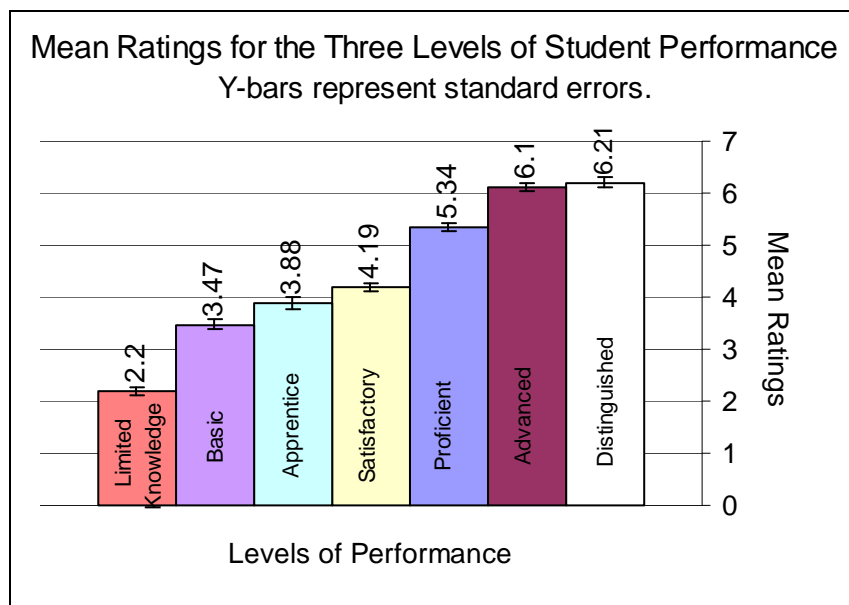
Concept	Mean	SD	Skew	Kurtosis	ICC
“Basic Level” N = 167					
Basic	3.47 ¹³	1.31	-1.85	4.67	.06
Apprentice	3.88 ²³	1.42	-.22	-.68	.04
Limited Knowledge	2.20 ¹²	1.12	.36	.39	-.02
“Proficient Level” N = 166					
Proficient	5.34 ⁴	1.08	-1.97	4.10	.01
Satisfactory	4.19 ⁴	0.82	1.56	2.86	.05
“Advanced Level” N = 167					
Advanced	6.10	1.01	1.38	3.37	-.06
Distinguished	6.21	1.20	-.28	-.10	.01

Note: Superscripts indicate significance at $p < .05$ in paired comparisons.

Table 4.6 No Mastery – Mastery Effect Sizes

Performance Category Comparison	Cohen's d
Apprentice- Limited Knowledge	1.31
Basic-Limited Knowledge	1.04
Apprentice- Basic	0.3
Proficient- Satisfactory	1.2

Figure 4.3 Mean Rating for No Mastery-Mastery Continuum.



As a reminder regarding the items measuring each concept on the no mastery-mastery continuum, concepts that were most positive or that indicated the highest level of mastery received a 7, the middle position on the scale was 4, and no mastery was assigned a 1. As indicated in Table 4.5, significant differences were found for four of the five planned comparisons.

RESEARCH QUESTION 3: RESULTS REGARDING TERMS WITH DEFINITIONS PROVIDED.

Research question 3 investigated the potential difference between performance category levels when definitions were provided. For research question 3, NAEP policy definitions for the performance levels *basic*, *proficient*, and *advanced* categories were provided to each respondent. Participants were separated into one of three groups. Group 1 was presented with the terms *limited knowledge*, *satisfactory*, and *advanced*. Group 2 was presented with the terms *apprentice*, *proficient*, and *distinguished*. Group 3 was presented with the terms *basic*, *proficient*, and *advanced*. Each group was presented with the same definitions for each of the three terms.

ANOVA was conducted in order to investigate these potential differences. As displayed in Table 4.7, results indicate that there is a significant mean difference in educators' perceptions of the terms that represent the two lowest levels of student performance (means are depicted in Figure 4.4). Specifically, significant difference in rating for those terms representing the "basic level" (i.e., *limited knowledge*, *apprentice*, and *basic*) where there is significant difference between *limited knowledge* and *apprentice* $t(155) = 4.96 p < .05$, and *limited knowledge* and *basic* $t(155) = 3.79 p < .05$. There was not a significant difference between *apprentice* and *basic* where $t(155) = 1.07 p > .05$. Overall the mean for *limited knowledge* was significantly less than *apprentice* and *basic*. The "proficient level" of student performance represented by the terms *satisfactory* and *proficient* also demonstrated a significant difference between terms $t(157) = 4.48 p < .05$, where *satisfactory* had a lower mean than *proficient*. Replicating the finding from research question 2, there was no significant difference between means for the "advanced level" of student performance after definitions were provided, that is, for the terms *advanced* and *distinguished* $t(156) = -1.97 p > .05$. In addition, when considering the

effect sizes from section 1 and section 2 of the online survey note that effect sizes were reduced on all accounts as displayed in Table 4.8.

Table 4.7 Group Ratings on the No Mastery-Mastery Continuum

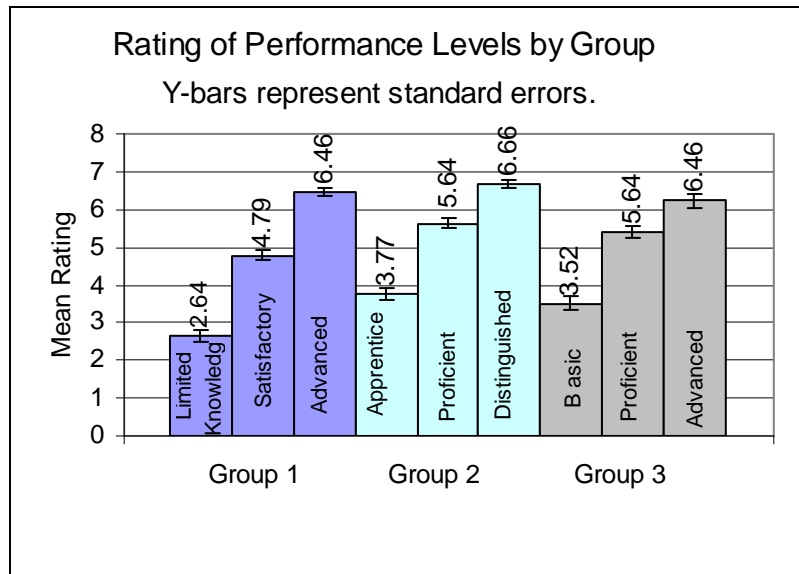
Levels	Terms	Group 1		Group 2			Terms	Group3		
		Mean	SD	Mean	SD	Mean		SD	N	
“Basic”	Limited Knowledge	2.64 ¹²	1.13	Apprentice	3.77 ¹	1.22	Basic	3.52 ²	1.23	157
“Proficient”	Satisfactory	4.79 ³	.93	Proficient ³	5.53	1.05	Proficient	5.53	1.05	159
“Advanced”	Advanced	6.35 ⁴	.99	Distinguished	6.66	.76	Advanced	6.35	.99	158

Note: Superscripts indicate paired comparison significance.

Table 4.8 Mastery Continuum Group Rating Effect Sizes

Performance Category Comparison	Cohen’s d
Apprentice- Limited Knowledge	0.96
Basic-Limited Knowledge	0.75
Apprentice- Basic	0.20
Proficient- Satisfactory	0.75

Figure 4.4 Performance Levels by Group



RESEARCH QUESTION 4: RESULTS REGARDING INTERVIEWS

Research question 4 was aimed at determining how performance category names were referred to during standard setting sessions. Four individuals were interviewed by phone, and each referenced a standard setting session they participated in within the last two years, so their detailed recollection may have been limited. Pseudonyms were created to protect the privacy of each interview participant, and the two tests referenced during the interviews were described as Test A and Test B.

While the first interview was conducted with Wendy, in reference to field-trial items for Test A, the last three interviews were conducted in reference to standard setting sessions for Test B.

The first two interview participants, Wendy and Marie, were described as technical staff. Technical staff was defined here as personnel that both observed the

standard setting process and conducted data analysis and other tasks as needed. The third interview was with Melissa, a “high-level observer” whose role during a standard setting session was strictly to oversee the process. The final interview was with Burton, a facilitator. Burton’s role during a standard setting session is typically to lead training and discussion. Marie, Melissa, and Burton all referenced Test B during their interviews. Please note that the interview participants did not set the standards themselves, as this was done by the standard setting participants, and will therefore offer a slightly different view point.

Two major themes emerged from the interviews: (a) variation in the construction of performance category definitions, and (b) variation in training and group discussion. Interview participant responses are outlined immediately below within each of these themes. Further discussion will outline some of the most common questions that surfaced during standard setting sessions, a brief discussion about the performance level categories and their definitions, followed by interviewee opinions about performance categories, and final summary.

VARIATION IN THE CONSTRUCTION OF PERFORMANCE CATEGORY DEFINITIONS

Performance category definitions took two forms. They were either predetermined and provided to the judges, or they were constructed by the judges during the standard setting session. The first variation was referenced by Wendy, the first interview participant, who noted that the performance category definitions were predetermined and simply provided to the judges during the standard setting session. In contrast, Marie, Melissa, and Burton in reference to Test B noted the second variation in which judges create the performance category definitions. Marie noted that although general definitions were provided for each performance level category, judges composed detailed, test specific, definitions of each performance level category as part of their training.

Additionally, Melissa, the observer, after reiterating that the definitions were constructed by the judges, also added that when the standard setting panel convened, standard setting was initiated with the level that was most important—the level equivalent to passing- *proficient*. Starting at this level, according to Melissa, was important in helping the judges make distinctions in order to define the remaining performance categories. Melissa also noted that discussion of definitions typically went from very general to very specific. Discussion would begin first with the broad definitions of the performance categories, moving the focus to the specific descriptions of the performance categories, and finally a discussion that focused on the relationship of the test items to each performance category. Furthermore, Melissa relayed that the performance level categories and their definitions were displayed on the wall or were projected on the overhead as the construction of the definitions and standard setting took place.

The facilitator, Burton, offered another perspective. In his opinion, when the definitions are already established (e.g., when states are anchoring new standards onto old standards) this makes it more difficult for the judges. He contends that “because judges did not define the terms themselves they do not have the buy-in that the group needs.” Further, Burton offered that providing definitions for judges may not be as satisfying. When group members are not involved in the process of developing the definitions, the same sense of ownership is not developed.

In summary, the three interview participants that referenced standard setting for Test B noted that as the definitions were constructed as a group that this led to judges who had ownership of the definitions they created. In contrast, for Test A, referenced by Wendy, the definitions were predetermined and simply provided to each judge. These two variations of defining performance categories speaks to the judges construction of

meaning, where in one sense the construction of meaning was literal, and in the other judges were asked to accept and internalize previously constructed meaning.

Related to the development of the category definitions is the participant understanding and internalization of the definition. It became apparent during the interviews that Test A, which used performance categories *basic*, *proficient*, and *advanced*, fostered more confusion among judges related to the prior knowledge associated with these terms. Wendy clarified this point as she offered that most participants relied on prior knowledge that *proficient* student performance was the target level (in reference to her standard setting session) when in fact, for that particular standard setting session, the *basic* level was considered passing. On aside, one might also note that percentage wise, the *basic* level of student performance on the NAEP is often more comparable to the *proficient* level of student performance on individual state test. It seems that... Wendy further reported that “most of the confusion stems strictly from the performance level categories. If *proficient* was not used for multiple purposes, then there probably would not be such a problem.” Marie’s perspective on performance categories and related questions was slightly different from Wendy’s perspective. For the standard setting session Marie participated in, the categories were not those commonly used across states and therefore Marie had a slightly different perspective to offer. Specifically, she believed that the differing opinions she observed about the connotation of the categories were more apparent at the beginning of the training. However, over the course of five days and through interactions with group members it was thought that the judges walked away with a more equal understanding or more equal connotation of each of the performance categories. In both cases, for Test A and Test B, Wendy and Marie made reference to a period in which clarification of meaning was sought by the judges. For Test A, it seemed that confusion stemmed from misconceptions about performance category

meaning due to their common use, and for Test B questions stemmed from a need to clarify what the categories meant to the judges connotatively.

Burton, the facilitator, was able to offer a bit more detail regarding the development of the performance categories themselves. It seems that this process (that of choosing the performance category), while not formally practiced in many states, is one that originated approximately 12 years ago. Burton had prepared to set standards in one state and as customary to his expectations, the state was asked to provide the number of levels and the names of each performance category before his work could begin. Interestingly, it turned out that this particular state considered the decision on what performance level categories to use as a crucial step, as a result, this state formed a separate committee composed of “high-level” individuals such as members of the state legislature, president of the state school board, dean of the community colleges, and president of the teacher association. Burton said that it was important to have this level of support in order to establish buy-in throughout the state. It was here that the idea of convening a separate group of people to choose the performance categories originated for Burton, and since this time he has been involved in at least 13 different state standard setting sessions. However, since this initial process (12 years ago), Burton has only completed this process formally with one other state, and he admits that the more recent implementation of selecting the performance level categories was far more formal than its predecessor 12 years earlier. Additionally, this process of selecting the performance level categories was conducted internally with state agency personnel in one other state. However, Burton did not equate that process with those previously referenced. He contended that “state people” are not the type of people who should be involved in choosing performance categories he further explained that one does not get the same

“bang for your buck” or the same public relations when it is conducted solely at the state level.

VARIATION IN TRAINING AND GROUP DISCUSSION

Two distinct differences in judge deliberation style emerged across the two testing programs. Test A and Test B standard setting sessions were similar in that a large group of about 20 judges was used to set standards for each grade and each subject. Differences emerged in whether there was discussion and deliberation among smaller groups (e.g., four groups of five) or if deliberation was conducted as a large group, meaning the entire group of 20 judges. Specifically, the researcher was interested in the procedure implemented to make decisions among the judges in the standard setting panels. Wendy mentioned that discussion and judge deliberation among small groups was the deliberation style of choice during the standard setting session for Test A. Discussion in small groups included topics such as what the performance level categories meant and clarification questions about the standard setting process itself. Wendy also noted that clarification questions about the performance level categories continued throughout most of the standard setting process within these small groups. When clarification was needed, typically a facilitator was consulted to alleviate any misconceptions; however, Wendy emphasized that the small group was not required to come to a consensus. It was also noted by Wendy that differentiation between performance levels was difficult. It was discovered at one point, in a previous field trial, that some judges were setting standards in terms of “typical student” performance, while the intent of the facilitator was for the group to set standards with the “borderline student” in mind. That is, set standards for each category according to the performance of a student who just barely met the standard for that category. In short, Test A utilized small groups for judge deliberation.

Marie, Melissa, and Burton on the other hand made it clear that all training and discussion for standard setting on Test B was held across the full group, composed of about 20 people. Melissa, the observer, from the same statewide testing program, noted that all training by the facilitators and judge deliberation, took place in the large group setting, and that as much of the process was standardized as possible.

Burton, the facilitator, reported that his organization had a strong bias to ensure all discussions took place in the large group setting. He adamantly believed that, “Standard setting is a collective process that involves very little psychometrics and statistics. It is much more about interactions.” He further stated that this activity should be conducted very much like a jury deliberates. “In a jury when you are given instructions you do not go off with two or three other people and decide what it is you are suppose to do or what to think. Every discussion you have with a jury is done in an open room so that everyone is privy to the information.” Furthermore, Burton believes that the facilitator should hear all of the comments being made; this allows the facilitator to shape conversations and to minimize misconceptions. For clarification, Burton interjected “...one thing about juries is that in a real jury you have to come to a consensus and in standard setting I do not care if they all agree or not, but that they should have at least used the same process to draw their conclusions.” He further elaborated and said that in a jury there becomes a group pressure to agree, but that in standard setting while there may be some pressure from other judges to agree that it is not what he expects to happen. Consensus does not have to happen in standard setting. The variation in training and group discussion across Test A and Test B are classical variations across methods used among states.

COMMON PARTICIPANT QUESTIONS

Across the four interviews it was clear that the most common questions during the typical standard setting session was concerning the process itself. Wendy reiterated this point “There were lots of clarification questions about the process and the materials that were given to them, and there also was discussion about when they should refer back to the categories and about how to interpret each of them.”

Marie also reported that the most common questions during her participation in the standard setting sessions for Test B related to the standard setting process itself. The more common questions according to Melissa, the observer, were more statements than they were questions. Melissa noticed and emphasized that participants were interested in voicing their own perspectives. Later she added that there were also questions about the standard setting process itself. Melissa clarified however, that the majority of the questions were voiced during group discussions as opposed to the initial training.

Burton recalled common questions in two areas: a) questions related to the internalizing of the task, and b) questions related directly to the performance level categories. “Say the term mastery is the highest level...one of the most important steps in standard setting is when you take the term mastery and make it more concrete for that panel, this aspect of standard setting typically takes one to one and a half hours. Mastery is a very generic concept. Everyone has their own perspective of what that means... so they spend an hour defining what each of those words mean and make it more concrete. People are always coming back around and asking questions like, ‘Why would a mastery student have to do this?’ At this point you have to remind them to go back to the definitions they wrote and for them to consider what they think the performance category means for their content area.”

PERFORMANCE LEVEL CATEGORIES AND THEIR DEFINITIONS

Interview participants were also asked whether the judges referenced the performance level categories or the category definitions more often. Interview participants from both Test A and Test B agreed that judges referenced the performance level categories and their definitions equally.

According to Wendy, because the performance level categories and their definitions were provided on sheets of paper in front of the judges during the standard setting for Test A, the two were so intertwined that the judges had to have referenced both. Additionally, judges were explicitly instructed not to reference their own personal students (if they were a teacher) but to think of a more representative group of students for each category and to focus on the “borderline” test takers or those test takers that just barely met each of the performance levels.

Marie believed that the performance categories were tied so strongly to their definitions that they reinforced their operational definitions, and that one was not referenced more than the other was. For clarification, it was noted that the definitions were displayed throughout the room on easel boards, and at times on the overhead when setting the actual standards.

Melissa also adamantly stated that the only way to understand what a person is suppose “to know and be able to do” is through the definitions, and that judges could not complete the task without taking the definitions into account.

Burton added that early on in the first round most judges frequently referred to the definitions displayed on easel boards around the room (the concrete descriptions and definitions) later to some extent, judges internalized what it meant to them and referred less to the definition displayed.

INTERVIEW PARTICIPANT OPINIONS ABOUT PERFORMANCE CATEGORIES

At the conclusion of the three interviews regarding Test B, participants were asked to describe their feelings on the use of labels or performance level categories in standard setting.

Marie relayed that providing categories is necessary. “If instead we used labels 1, 2, and 3 as performance level categories that would not be sufficient.” Marie emphasized that comparing across states would be inherently problematic since their discussions and standards were not made in concert. “The different labels used across states are only problematic when you start comparing across states. What one state thinks is important for students to know another state might not think is important for students to know.”

Melissa’s opinion on performance categories is that standard setting “is very political and that the terms carry with them lots of baggage that may or may not reflect the skills noted at that level of performance. In some states, *proficient* is used to describe basic skills. For example, in one definition, *proficient* might refer to beginning performance at the 4th grade level and in another state it might refer to higher level 3rd grade work, yet each is called *proficient*. Overall, the labels get in the way. Any terms that suggest failure generally are not favored. People tend to look for terms like *nearing the standard*, some states just use terms like *level 1*, *level 2*, and *level 3*.”

Burton asserted that “Once the labels are chosen to a certain extent you have already predetermined the percent of kids who are in those cells. If you use a term like *advanced*, *exceptional*, *superior*, or *mastery*... all of those have different connotations and I think they imply a different point on the scale.” Furthermore, he adds, “common words like *basic*, *proficient*, and *advanced* are a problem, as a result the definition of *proficient* varies significantly across states.” In addition, Burton believes that the process

of choosing labels should be standardized. “The determination of that label is a critical aspect of the entire enterprise.”

SUMMARY

In summary, the first inquiry was to determine whether teachers, principals, and administrators perceived a statistically significant difference across connotations of the seven performance categories. Connotative differences existed across all levels of performance category terms, based upon the online survey responses. Specifically, among the terms representing the “basic level” of student performance *apprentice* had the most positive connotation, followed by *basic*, and *limited knowledge*. For the *proficient* level of student performance *proficient* had a more positive connotation than *satisfactory*, and for the *advanced* level of student performance *distinguished* had a more positive connotation than *advanced*.

The second inquiry was to determine whether a statistically significant difference existed among educators’ perceptions of the seven performance categories when presented on a mastery continuum. There was a significant difference in the perception of mastery overall for the terms representing the *basic* level of student performance and the terms representing the *proficient* level of student performance. When comparing the terms by category, it appears that there is a significant difference between the terms *apprentice* and *limited knowledge*, and *basic* and *limited knowledge*, and *apprentice* and *basic* as they are perceived to convey mastery. Teachers, principals, and administrators also perceived a difference in the performance level categories *satisfactory* and *proficient* each of which are purported to represent the middle, or the *proficient* level, of student performance. Finally, the terms representing the third and highest performance level, *advanced* and *distinguished* were not determined to be significantly different in terms of mastery by educators and administrators.

The third concern of inquiry was to determine whether a statistically significant difference existed across three groups' perception of the performance level categories when definitions for the categories were provided. Results suggest that after definitions were provided, in an effort to provide the same context or the same denotative meaning for performance level categories, *limited knowledge* was still seen as the most negative of the three "basic level" terms. However, after providing the definitions, *apprentice* and *basic* were no longer significantly different from each other on the mastery continuum. An investigation of the second level of performance suggests that after taking into account definitions, that the terms *proficient* and *satisfactory* are also still significantly different, where *proficient* is seen as describing a student with more mastery than a student deemed *satisfactory*. Finally, no significant difference was shown between the perception of the terms *advanced* and *distinguished*. Here it seems that only terms that were very different connotatively, as demonstrated through the large effect sizes, maintained their significant differences after definitions were provided.

The fourth and final inquiry was to determine how performance category names were referenced during a typical standard setting session. For each interview, the distinction occurred not in how the performance categories were referenced but how their definitions were created and referenced. It seems that the major difference in understanding and communication of terms and definitions in standard setting processes hinged on the involvement of the judges or participants in the development of the definitions. For the three interview participants that referenced Test B, each articulated their belief that judges understood and at times internalized the definitions they had developed. There was a sense of ownership. Wendy, who referenced Test A, articulated that question and concern about the meaning of the terms continued throughout the entire

standard setting process. This was the main difference between the two standard setting sessions that were referenced in regards to the use of category names.

CHAPTER 5: SUMMARY, IMPLICATIONS, LIMITATIONS, CONCLUSION

NCLB and its mandates for states brought forth the possible importance of connotative meaning of performance categories used in standard setting. If all states are to report to the U.S. Department of Education the percent of students performing at the *proficient* level, with the expectation that 100 percent of students meet the *proficient* mark, it may be important to ensure that the *proficient* mark across states are comparable. However, there are clearly two perspectives concerning the significance of this issue. First, researchers argue that NCLB does not mandate equivalence across states but instead mandates improvement within each state. In short, definitions of *proficient* need not be equivalent or compared. On the other hand, some researchers argue, with the standardization put forth by NCLB, comparison across states is inevitable. NCLB has mandated both adequate yearly progress (AYP) and that 100 percent of students meet the proficient mark as each are separate goals, one as a means to the end. Consequently, as the U.S. Department of Education seeks to hold states accountable for student performance many states have passed the same accountability onto their students. For many students their ability to reach the *proficient* mark not only reflects on their state, but also will determine whether they will move to the next grade or graduate high school. The high stakes associated with the increase in testing via NCLB speaks to the importance of this study.

The major objectives of this study were to determine if performance categories commonly used in standard setting hold connotative differences, and to investigate how these performance categories were referenced in standard setting sessions. This research was framed by the theoretical perspectives of Osgood, Loftus, Kintsch, and Bakhtin.

Specifically, these frameworks guided the assessment of possible differences in connotative meaning across various performance categories as used in standard setting. In the following sections, a review of the results, an integrative summary of findings, a discussion of implications for practice, study limitations, future directions, and a final conclusion are discussed.

OVERVIEW OF RESULTS

Overall, significant mean differences were found across the seven performance categories. Consistent differences across three terms representing the *basic* level of student performance (i.e., *limited knowledge*, *apprentice*, and *basic*) and the two terms representing the *proficient* level of student performance (i.e., *satisfactory* & *proficient*) were found. One exception surfaced when definitions were provided, after which the mean difference between *apprentice* and *basic* was no longer significant. Additionally, it was only on the evaluative factor for the semantic differential that a significant mean difference between the terms representing the *advanced* level of student performance (i.e., *advanced* & *distinguished*) existed. In short, educators across the nation perceive many of the terms used to set standards to have different meanings.

Educators rated the terms on three scales, a) on a master scale, b) connotatively, and c) on a mastery scale based on definitions. For instance, when participants were provided with the performance category only, a difference in the perceived level of mastery occurred for several categories. All of the terms at the basic level of student performance had significantly different mean ratings from each other (i.e., *basic*, *limited knowledge*, and *apprentice*). The effect size related to these mean differences emphasizes the practical significance of the findings. The effect size as relayed in Table 4.6 indicates the large difference in the level of mastery between *apprentice* and *limited knowledge*, *basic* and *limited knowledge*, and a more moderate difference between *apprentice* and

basic. Given the large effect size cited for each category when compared to *limited knowledge* on the mastery continuum, it could be said that educators view limited knowledge student as students with lower levels of mastery. The overall perceived level of mastery was much lower for the *limited knowledge* category, and there was a relatively small difference between the *apprentice* and *basic* categories. In addition, educators reported a large difference with respect to their perception of mastery for the categories *proficient* and *satisfactory*, with *proficient* being perceived as a much higher level of mastery.

When educators evaluated the terms on the semantic differential their perceptions were fairly consistent. Mean differences on the evaluative scale for the semantic differential were also found across all performance categories. The majority of the evaluative mean differences were rather large in terms of effect sizes. In particular, the difference in means for the terms at the basic level (i.e., *basic*, *limited knowledge*, and *apprentice*) ranged from .88 to 1.49. The difference in means for the categories *basic* and *apprentice* had a medium effect size. While a significant difference in the mean evaluative rating for the *advanced* and *distinguished* categories was found, the effect size was rather small as displayed in Table 4.4. Notice for these measures of evaluation, that is, whether the performance categories were seen as good or bad, the effect size is similar in magnitude to those found on the mastery continuum discussed previously hence supporting educators' perceived difference across performance categories.

Conversely, when definitions were provided with each of the performance categories, differences persisted on the mastery continuum for only three of the four comparisons. Mean difference ratings for the categories *apprentice* and *limited knowledge*, *basic* and *limited knowledge*, and *proficient* and *satisfactory* each had large effect sizes. The difference in means for *apprentice* and *basic* did not persist after

definitions were provided. Also, note the change in effect size estimates from the previous analyses as displayed in Tables 4.6, 4.4, and 4.8. While the differences remained for some of the categories, when definitions were provided the effect sizes appear to be reduced which suggests that the definitions, although limited to a few lines, may have helped to mitigate the perceived differences in the level of mastery between performance categories. Interview participants concurred that while individual differences in performance category meaning proliferated in the beginning, after training and instructions performance category meaning became more of a consensus.

INTEGRATIVE SUMMARY

The empirical results from the online survey coupled with the exploratory results from the four interviews led to several suggestions. These suggestions relate to the appropriateness of comparisons made across states, the need for a closer look at the deliberation style across standard setting techniques, and further study of the relationship of connotative and denotative meaning of performance categories.

Comparisons across States

To begin, the results of this study suggest that if comparisons of student performance across states are made some students may be “left behind.” For illustration, consider a state using the terms *limited knowledge*, *satisfactory*, and *advanced* to describe student performance, in contrast to a second state using the terms *limited knowledge*, *proficient*, and *advanced*. If the differences in connotations of the terms *satisfactory* and *proficient* persist in the context of standard setting and if these differences extend to eventual cutscore placements, then the results of this study support that the second state (the state using the term *proficient*) might possibly designate fewer students in the *proficient* level. This example is closely related to the work of Loftus and Palmer (1974)

in which they reported that the verb *smashed* elicited higher estimates of speed than questions that used alternate verbs such as *collided*, *bumped*, *contacted*, or *hit*. Likewise, for this study *proficient* consistently elicited higher mean ratings on the mastery continuum indicating the perception that typical *proficient* students know more than typical *satisfactory* students do. Additionally, the mean scale scores on the evaluative factor indicate that *proficient* had a more positive connotation than *satisfactory*. Not only was the mean difference between *proficient* and *satisfactory* significantly different, but the practical significance is supported by the large effect size. Similarly, the same comparison of differences exists between *limited knowledge*, *apprentice*, and *basic* where *limited knowledge* was consistently rated lower and had a less favorable connotation than both *apprentice* and *basic*. The common goal across states to have 100 percent of students at the *proficient* level may still leave some students behind.

In short, these data suggest that some performance categories elicit perceptions of differential levels of knowledge; as a result, if these differences continue through the standard setting sessions and penetrate to the cutscores, it is possible that states could indirectly limit the percent of students who obtain the level of *proficient*. Burton, the facilitator, also hinted at this potential implication, “Once the labels are determined, to a certain extent you have already predetermined the percent of kids who are in those cells.” The significance of connotative meaning in the context of standard setting might be important if comparisons across states are expected. Conversely, it also appears from the data that differences in connotations may not persist if enough definition and discussion are provided.

Deliberation Style

Second, the deliberation style referenced for Tests A and B, and the standard setting techniques used, is important when considering the potential impact of

connotation on standard setting. The interview summaries represent two forms of training and deliberation style in standard setting. In one form discussed here, the goal of the facilitator was to keep most if not all discussion in a large group format. In the second deliberation form, as was referenced in regards to Test A, small group discussion and deliberation was utilized in addition to the large group format. Given the marked differences in these deliberation styles, Bakhtin (1984) might suggest that because each setting provides different experiences and interactions between the judges it could potentially lead to different conclusions and possibly different standards. Bakhtin's perspective would suggest that since our experiences give way to the construction of meaning and that meaning is not embedded in words themselves, but is constructed between people that the two deliberation forms are not equivalent. The form of meaning negotiation and the setting (that is large versus small group) is quite dissimilar across the two deliberation styles. According to Bakhtin, the meaning making process is continuously acting and reacting. This interpretation suggests that the meaning arrived at by judges who participate in the small group versus those who participate in large group discussions could potentially be very different and therefore result in disparate standards.

The implications for judges reaching conclusions in small versus large groups should not be informed solely by the data here. Yet, psycholinguistic theory supports the notion that deliberation that takes place in four small groups is likely to have more variability than deliberation conducted in one large group. In fact, according to comments made during the interviews, it seems those judges who were trained and deliberated in the large group setting and literally constructed meaning together, appeared to have fewer prolonging questions concerning the meaning of each performance category. However as Burton noted, after a couple of rounds of deliberation (there are typically three depending on the method) judges seemed to internalize the performance category definitions and

made their judgments with less reference to the definitions provided for them throughout the room.

Connotative and Denotative Meaning

Regarding connotative and denotative meaning, the data here suggest that after providing definitions for the performance categories a significant difference in mean ratings persisted. However, while the existence of differences across categories was maintained (except for the *apprentice* vs. *basic* pair), the effect size of those differences was lessened. Specifically, the effect sizes of the mean difference between the *basic* level terms were reduced from 1.31 to .96 for the *apprentice-limited knowledge* pair and from 1.042 to .75 for the *basic-limited knowledge* pair. Given the continued significant mean difference across performance categories suggests that connotative meaning might at times supersede denotative meaning. For example, respondents reported a lower mean rating on the mastery continuum for student performance at the *limited knowledge* level than what was reported for student performance at the *apprentice* or *basic* level. In addition, for the second level of student performance, survey respondents perceived *satisfactory* student performance as significantly lower than *proficient* student performance on the mastery continuum. Yet, as demonstrated in Table 2.1, these terms are used to denote the same level of performance across at least three states Arkansas, Kentucky, and Oklahoma, and each category was defined in exactly the same manner for this study. To sum up, the data show simply because we define and designate terms to represent a category does not mean they have the same connotative meaning. Yet, it must also be noted that the differences in the category terms diminished after definitions were provided. Therefore it could be hypothesized that further more detailed definitions could diminish the differences altogether.

Support for the existence of connotative differences emerged during the phone interviews as well. Interview participants noted confusion and difficulty when commonly used performance categories (e.g., *basic*, *proficient*, and *advanced*) were not used consistently across states and referred to incomparable levels of performance. For example, in Mississippi, *proficient* denotes passing but in Missouri, *proficient* denotes the highest level of student achievement. According to Perrig and Kintsch (1985) prior knowledge is integral to achieving understanding; therefore, the fact that the category *proficient* is so commonly associated with NAEP, and more often considered the passing level across states, relates to the influence of prior knowledge in the context of standard setting. The prior knowledge associated with commonly used performance categories may hinder understanding and may be another matter for consideration in further standard setting research.

In conclusion, overall findings support that educators and administrators perceive many of the terms that are commonly used in standard setting and testing as connotatively different. While the definitions that are provided during standard setting offer a context and a common ground, it is not conclusive that definitions eliminate individual perceptions or connotative meaning. During the interviews, the consensus was that performance categories are intertwined with their definitions such that judges mostly rely on the definitions when setting standards. Still, the results here suggest that at times judges in standard setting sessions could possibly rely on connotative meaning more than denotative meaning of performance categories when the terms used are connotatively potent. In addition, as states strive to meet the expectations of NCLB it is apparent that the potential for connotative differences between terms used across performance levels could have greater impact. The differences found here support suggestions for future

research into the implications of connotation on the placement of cutpoints in standard setting.

IMPLICATIONS FOR PRACTICE

Implications for practice based on study findings imply that connotative meaning at times over-shadows that of denotative meaning. It should be noted that, because the survey was administered online, there is little assurance that respondents read the instructions and definitions. Some might argue that the connotative meaning of a word or performance category is not an issue in standard setting because performance categories are operationalized, and judges are trained to understand and internalize the meaning of each performance category. Nonetheless, the conclusions here support a claim asserted by the facilitator during the interviews, “Just because you call them both *proficient* does not make them equivalent. There is no way that we could possibly get the same percentage of students in a *satisfactory* level as we do in a *mastery* level.” The results of this study are exploratory and the potential impact for the standard setting process varies depending on the technique and implementation. Implications for practice and future research will also vary depending upon whether or not there are iterations of decisions, whether or not judges are given impact data, among many other potential variations in practice. In short, the results presented here are preliminary and need further investigation, and the potential impact will vary.

It could be argued that standardization of the standard setting process is critical to comparability. While it is not practical to enforce national content standards, national performance standards, or to mandate that states all use the same standard setting techniques, some would argue that efforts should be made to make as many aspects of the standard setting process routine and comparable as possible. While the current study focused on the two most used techniques across the states (i.e., the Bookmark and the

modified-Angoff techniques), the difference in techniques also relates to the impact or the implications for practice. Comparable standard setting procedures would better lend its outcome to meeting the overarching goal of No Child Left Behind. Decisions on the terms used to represent each performance category are a vital step to every implementation of standard setting. If the goal of the U.S. Department of Education is to ensure that all students are achieving at the *proficient* level, actions to investigate the connotative similarity of one state's level of *proficient* to another are suggested. Although peer reviews took place to ensure that state standards were rigorous and aligned with state content standards, the connotations of the terms used to describe student performance might themselves be "leaving some children behind." It is suggested that future research investigate the implications of connotation on the standard setting procedures and outcomes.

STUDY LIMITATIONS

As with all research, this study too had limitations. The abstract nature of the study (that is the absence of terms associated directly to a grade or subject) and the mode of delivery were both an advantage and disadvantage. Below I discuss the trade-offs between the mode of survey delivery, the nature of the survey, survey instruments, and interview and pre-testing participants.

To begin, while administering an online survey facilitated the overall process it also produced some expected and unexpected limitations. The online survey aided in contacting a wide range of teachers and administrators over a short period, and facilitated efficient data collection and data analysis. On the other hand, administering an online survey restricted the potential respondents to educators and administrators who had access to the Internet. While limited access to the Internet was not believed to be a factor for this study, intermittent local Internet problems during survey administration could

have limited survey responses. Another limitation that emerged related to the mode of survey delivery was broken hyperlinks. It was discovered through phone calls that the email survey announcement included a handful of broken survey hyperlinks. Broken links were the result of lines that wrapped within email programs for some survey respondents; this resulted in hyperlinks that when clicked did not launch the survey. Broken links appeared to be a random problem and was likely related to individual email settings (according to the programmer). It was resolved for some respondents over the phone, but it is possible that it remained an impediment for others. Additionally, related to the mode of the survey it should be noted that online surveys could serve as limitations in schools as many schools have strict filters and will not allow receipt of unexpected email messages. Also, as is common with the nature of electronic information some teacher, principal, and superintendent email addresses posted on websites were out of date. In the end, limitations associated with administering an online survey were mitigated using postcard follow-ups.

The second area of limitations related to the abstract nature of the study. Although definitions, instructions, and other prompts were offered as context for the survey, the terms rated by survey respondents were not specific to any subject or grade level but were about students in general. Moreover, I can not be sure that survey respondents were operating from the same perspectives. That is, it can not be determined that respondents truly read the definitions when provided, so this too limits the generalizability of the results. Additionally, the performance category definitions that were provided were brief (one to two sentences) in comparison to more detailed lengthy definitions typically provided during standard setting sessions.

A third limitation relates to the creation of the mastery continuum. This tool has not been previously used for this type of measurement (perception of student knowledge

level) and, although pre-tested, was not psychometrically validated. It is possible that the seven point continuum limited the potential distinction of the higher-level terms *advanced* and *distinguished*. It is proposed here that future studies extend the scale to around nine points to allow for a finer measurement. The semantic differential coupled with the nature of the study, and the ability to devise accurate adjective-pair items to represent the potency and activity dimensions served as another potential limitation. Directions for the semantic differential were difficult to make brief and concise. At first, Osgood's original semantic differential instructions were utilized for the online survey. During pre-testing teachers sighed with exasperation at the lengthy instructions. Teacher feedback led to the final directions, yet further refinement and pre-testing of the instructions is suggested.

Finally, the sample of interview and pre-testing participants was rather small and offered a relatively limited perspective. Phone interview participants each referenced tests that utilized the Bookmark (sometimes referenced as item-mapping) procedures. While three of the respondents referenced the item mapping process for Test B, only one of the respondents provided a perspective for Test A procedures. Also, due to the role of interview participants during standard setting a limited perspective was provided which limits the generalizability of the conclusions drawn. A missing perspective is that of the judge, the person responsible for placement of a cutpoint or standard. Interviewing judges might offer a fresh point of view though it is difficult to obtain their participation given the promises of nondisclosure typically signed during standard setting sessions. Furthermore, the number of participants used to pretest the online survey and interview questions was rather small and might be seen as a limitation to the study. Additionally, pre-testing participants were recycled and after a couple of rounds of viewing the survey, they may have become too familiar with the expectations of the survey and researcher.

FUTURE DIRECTIONS

The following section introduces future directions for the current study. Suggested studies should be considered fertile ground for continued research:

1. A similar study might investigate group dynamics from a psycholinguistic viewpoint. A study of this nature might be more effective if completed in an actual standard setting session in which there is a comparison between standard setting sessions that use small group discussion vs. large group discussion. Empirical evidence is needed to incorporate the psycholinguistic theories of meaning and group dynamics.
2. Different performance level categories could be used in addition to the seven studied here.
3. Researchers might consider having subjects write the definition of the categories as each is coupled with connotatively different terms. For example, one might have a group of teachers write the definition for *basic*, *proficient*, and *advanced* compared to *failure*, *proficient*, and *advanced*. This variation might provide a better idea of how the context of the surrounding words affects the perception of mastery for each category.
4. It might also be of interest to increase scales to about nine points to allow for more differentiation. A finer distinction on the continuum might allow survey respondents to make an enhanced rating of terms.

CONCLUSION

Using respondent ratings of performance categories on an online survey and through conducting interviews, it was found that performance categories that are denotatively similar might be connotatively quite different. The potential implications in standard setting and consequent implications for the goals of No Child Left Behind call

for further study of connotation in standard setting. I would like to close with a quote cited previously in chapter 2, “the names we give things, events, and people determine our behavior towards them” (Burke, 1965, p. xiv). Likewise Burton, the facilitator, stated “the names of the performance categories, to some extent predetermine how many students will be put in that category.” What we say and how we say it matters, especially in the context of standard setting.

Appendices

Appendix A: Online Survey

WELCOME!

Performance Categories Used in Setting Standards

The purpose of this survey is to attempt to measure the meanings of certain performance labels by having you judge these labels against a series of descriptive scales. The survey is composed of four sections, a total of 12 screens, and should take you less than ten minutes to complete. Please note that participation in this research study is very important as it may stimulate change in the field of testing, and will provide me with information to better inform the standard setting process in your state. You will be notified via email when the survey results are available. Confidentiality of your responses will be ensured by removing all personal identifiers upon completion of the survey, and results will only be reported in summary form.

[Begin Survey](#)

Please note that no guarantees can be made regarding the interception of data sent via the Internet by any third parties.

Appendix A Continued: Online Survey

SECTION 1

Instructions: Please select a radio button along the "No Mastery" to "Mastery" continuum for each category. The continuum scale reflects typical student performance. Please make your choices based on what the definition of each category name means to you.

Advanced



No Mastery Mastery

Apprentice



No Mastery Mastery

Basic



No Mastery Mastery

Distinguished



No Mastery Mastery

Appendix A Continued: Online Survey

Distinguished

No Mastery Mastery

Limited Knowledge

No Mastery Mastery

Satisfactory

No Mastery Mastery

Proficient

No Mastery Mastery

[Next Page](#)

Appendix A Continued: Online Survey

SECTION 2

INSTRUCTIONS: In this section, there are a total of seven key words related to student performance. Each key word has nine descriptive scales below it. Select the radio button in the middle position if you consider the key word at the top of the page to be neutral on the scales. The remaining three buttons on either side represent the strength of relationship between the key word at the top of the page and each of the descriptive scales. The closer the button is to the end of the scale the stronger the relationship. *Please remember that your selection should reflect what each key word means to you.*

Proficient

Good	<input type="radio"/>	<input checked="" type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	Bad
Pleasant	<input type="radio"/>	<input checked="" type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	Unpleasant
Valuable	<input type="radio"/>	<input checked="" type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	Worthless
Weak	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input checked="" type="radio"/>	<input type="radio"/>	Strong
Small	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input checked="" type="radio"/>	<input type="radio"/>	Large
Light	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input checked="" type="radio"/>	<input type="radio"/>	Heavy
Sharp	<input type="radio"/>	<input checked="" type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	Dull
Active	<input type="radio"/>	<input checked="" type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	Passive
Fast	<input type="radio"/>	<input checked="" type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	Slow

Next Page

Appendix A Continued: Online Survey

SECTION 2

INSTRUCTIONS: In this section, there are a total of seven key words related to student performance. Each key word has nine descriptive scales below it. Select the radio button in the middle position if you consider the key word at the top of the page to be neutral on the scales. The remaining three buttons on either side represent the strength of relationship between the key word at the top of the page and each of the descriptive scales. The closer the button is to the end of the scale the stronger the relationship. *Please remember that your selection should reflect what each key word means to you.*

Distinguished

Good	<input checked="" type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	Bad
Pleasant	<input checked="" type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	Unpleasant
Valuable	<input checked="" type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	Worthless
Weak	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input checked="" type="radio"/>	<input type="radio"/>	Strong
Small	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input checked="" type="radio"/>	<input type="radio"/>	Large
Light	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input checked="" type="radio"/>	<input type="radio"/>	Heavy
Sharp	<input type="radio"/>	<input checked="" type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	Dull
Active	<input type="radio"/>	<input checked="" type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	Passive
Fast	<input type="radio"/>	<input checked="" type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	Slow

Next Page

Appendix A Continued: Online Survey

SECTION 2

INSTRUCTIONS: In this section, there are a total of seven key words related to student performance. Each key word has nine descriptive scales below it. Select the radio button in the middle position if you consider the key word at the top of the page to be neutral on the scales. The remaining three buttons on either side represent the strength of relationship between the key word at the top of the page and each of the descriptive scales. The closer the button is to the end of the scale the stronger the relationship. *Please remember that your selection should reflect what each key word means to you.*

Basic

Good	<input type="radio"/>	<input type="radio"/>	<input checked="" type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	Bad
Pleasant	<input type="radio"/>	<input type="radio"/>	<input checked="" type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	Unpleasant
Valuable	<input type="radio"/>	<input type="radio"/>	<input checked="" type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	Worthless
Weak	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input checked="" type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	Strong
Small	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input checked="" type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	Large
Light	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input checked="" type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	Heavy
Sharp	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input checked="" type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	Dull
Active	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input checked="" type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	Passive
Fast	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input checked="" type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	Slow

Next Page

Appendix A Continued: Online Survey

SECTION 2

INSTRUCTIONS: In this section, there are a total of seven key words related to student performance. Each key word has nine descriptive scales below it. Select the radio button in the middle position if you consider the key word at the top of the page to be neutral on the scales. The remaining three buttons on either side represent the strength of relationship between the key word at the top of the page and each of the descriptive scales. The closer the button is to the end of the scale the stronger the relationship. *Please remember that your selection should reflect what each key word means to you.*

Limited Knowledge

Good	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input checked="" type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	Bad
Heaven	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input checked="" type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	Hell
Happy	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input checked="" type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	Unhappy
Weak	<input type="radio"/>	<input checked="" type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	Strong
Small	<input type="radio"/>	<input checked="" type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	Large
Light	<input type="radio"/>	<input checked="" type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	Heavy
Scary	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input checked="" type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	Not
Active	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input checked="" type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	Passive
Fast	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input checked="" type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	Slow

Appendix A Continued: Online Survey

SECTION 2

INSTRUCTIONS: In this section, there are a total of seven key words related to student performance. Each key word has nine descriptive scales below it. Select the radio button in the middle position if you consider the key word at the top of the page to be neutral on the scales. The remaining three buttons on either side represent the strength of relationship between the key word at the top of the page and each of the descriptive scales. The closer the button is to the end of the scale the stronger the relationship. *Please remember that your selection should reflect what each key word means to you.*

Satisfactory

Good	<input type="radio"/>	<input type="radio"/>	<input checked="" type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	Bad
Pleasant	<input type="radio"/>	<input type="radio"/>	<input checked="" type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	Unpleasant
Valuable	<input type="radio"/>	<input type="radio"/>	<input checked="" type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	Worthless
Weak	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input checked="" type="radio"/>	<input type="radio"/>	<input type="radio"/>	Strong
Small	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input checked="" type="radio"/>	<input type="radio"/>	<input type="radio"/>	Large
Light	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input checked="" type="radio"/>	<input type="radio"/>	<input type="radio"/>	Heavy
Sharp	<input type="radio"/>	<input type="radio"/>	<input checked="" type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	Dull
Active	<input type="radio"/>	<input type="radio"/>	<input checked="" type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	Passive
Fast	<input type="radio"/>	<input type="radio"/>	<input checked="" type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	Slow

Next Page

Appendix A Continued: Online Survey

SECTION 2

INSTRUCTIONS: In this section, there are a total of seven key words related to student performance. Each key word has nine descriptive scales below it. Select the radio button in the middle position if you consider the key word at the top of the page to be neutral on the scales. The remaining three buttons on either side represent the strength of relationship between the key word at the top of the page and each of the descriptive scales. The closer the button is to the end of the scale the stronger the relationship. *Please remember that your selection should reflect what each key word means to you.*

Advanced

Good	<input type="radio"/>	<input checked="" type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	Bad
Pleasant	<input type="radio"/>	<input checked="" type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	Unpleasant
Valuable	<input type="radio"/>	<input checked="" type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	Worthless
Weak	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input checked="" type="radio"/>	Strong
Small	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input checked="" type="radio"/>	Large
Light	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input checked="" type="radio"/>	Heavy
Sharp	<input type="radio"/>	<input checked="" type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	Dull
Active	<input type="radio"/>	<input checked="" type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	Passive
Fast	<input type="radio"/>	<input checked="" type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	Slow

Next Page

Appendix A Continued: Online Survey

SECTION 2

INSTRUCTIONS: In this section, there are a total of seven key words related to student performance. Each key word has nine descriptive scales below it. Select the radio button in the middle position if you consider the key word at the top of the page to be neutral on the scales. The remaining three buttons on either side represent the strength of relationship between the key word at the top of the page and each of the descriptive scales. The closer the button is to the end of the scale the stronger the relationship. *Please remember that your selection should reflect what each key word means to you.*

Advanced

Good	<input type="radio"/>	<input checked="" type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	Bad
Pleasant	<input type="radio"/>	<input checked="" type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	Unpleasant
Valuable	<input type="radio"/>	<input checked="" type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	Worthless
Weak	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input checked="" type="radio"/>	<input type="radio"/>	Strong
Small	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input checked="" type="radio"/>	<input type="radio"/>	Large
Light	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input checked="" type="radio"/>	<input type="radio"/>	Heavy
Sharp	<input type="radio"/>	<input checked="" type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	Dull
Active	<input type="radio"/>	<input checked="" type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	Passive
Fast	<input type="radio"/>	<input checked="" type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	Slow

Next Page

Appendix A Continued: Online Survey

Section 3 (Version 1)

SECTION 3

Instructions: Definitions for each student-performance category are provided below. Read the definition and then select a radio button along the continuum to reflect what typical student performance for each category WOULD MEAN TO YOU.

Apprentice

Definition: Partial mastery of prerequisite knowledge and skills that are fundamental for proficient work at each grade.



No Mastery Mastery

Proficient

Definition: Solid academic performance for each grade assessed. Students reaching this level have demonstrated competency over challenging subject matter, including subject-matter knowledge, application of such knowledge to real-world situations, and analytical skills appropriate to the subject matter.



No Mastery Mastery

Distinguished

Definition: Superior performance.



No Mastery Mastery

Next Page

Appendix A Continued: Online Survey


Section 3 (Version 2)

SECTION 3

Instructions: Definitions for each student-performance category are provided below. Read the definition and then select a radio button along the continuum to reflect typical student performance for each category.

Basic


Definition: Partial mastery of prerequisite knowledge and skills that are fundamental for proficient work at each grade.



No Mastery Mastery

Proficient

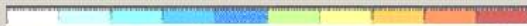
Definition: Solid academic performance for each grade assessed. Students reaching this level have demonstrated competency over challenging subject matter, including subject-matter knowledge, application of such knowledge to real-world situations, and analytical skills appropriate to the subject matter.



No Mastery Mastery

Advanced

Definition: Superior performance.



No Mastery Mastery

[Next Page](#)

Appendix A Continued: Online Survey


Section 3 (Version 3)

SECTION 3

Instructions: Definitions for each student-performance category are provided below. Read the definition and then select a radio button along the continuum to reflect what

Limited Knowledge

Definition: Partial mastery of prerequisite knowledge and skills that are fundamental for proficient work at each grade.



No Mastery Mastery

Satisfactory

Definition: Solid academic performance for each grade assessed. Students reaching this level have demonstrated competency over challenging subject matter, including subject-matter knowledge, application of such knowledge to real-world situations, and analytical skills appropriate to the subject matter.



No Mastery Mastery

Advanced

Definition: Superior performance.



No Mastery Mastery

Next Page

Appendix A Continued: Online Survey

SECTION 4

Some states use teachers to set the standards for their state's standardized test.
Have you participated in a standard-setting session like this before?

- Yes
 No

How many years have you been an educator in K-12 schools?

11 ▾

Please indicate which subject(s) you currently teach (check all that apply).

- English/Language Arts
 Mathematics
 Science
 Social Studies
 Other

Comments:

Appendix A Continued: Online Survey

THANK YOU FOR YOUR TIME!

You will be notified via email when the results are available. The raffle will take place at the conclusion of our celebration luncheon on Wednesday, October 26, 2011 at 5:00 PM.

Thank you again for your time and participation.

Sincerely,
Wendy Dutt

Appendix B: Pre-Notification Email

Dear [FirstName],

A few days from now you will receive an email request with the subject “University of Texas Online Survey.” Superintendent [CustomData] has granted me permission to survey teachers in your school.

The purpose of the survey is to determine the meanings of labels used to describe student performance on standardized tests.

I am writing in advance because we have found that many people like to know ahead of time that they will be contacted. Participation in this research study is very important as it may stimulate change in the field of testing AND may aid in better alignment of federal and state testing expectations.

It is only with the generous help of people like you that our research can be successful. As a way of saying thank you, I will conduct a raffle for three \$50 prizes to be awarded April 16, 2004. Thank you for your time and consideration.

Sincerely,
Winona M. Burt
Doctoral Candidate, Quantitative Methods
Educational Psychology
University of Texas
Austin, TX 78712
Office: 512-385-5520
Mobile: 512-731-5723

Appendix C: Survey Launch Email

Dear [FirstName],

I am conducting a survey as part of my Ph.D. requirements. The purpose of the survey is to determine the meanings of labels used to describe student performance on standardized tests across the United States.

Only a small sample of educators from 48 schools across the country has been asked to participate in the survey and your professional input is essential for the completion of this study. The survey is similar to an opinion survey, should take less than 10 minutes of your time, and will provide me with information to better inform the standard setting process in your state. Your responses to the survey will be held in the strictest confidence.

[LastName] granted me permission to contact teachers in your school and you were randomly selected.

If you have any questions at all, please do not hesitate to call at 512-385-5520 or email me wburt@mail.utexas.edu. As a way of saying thank you, I will conduct a raffle for three \$50 prizes. The first prize will be awarded April 2, 2004.

To begin the survey please click here:

[http://evalsoft07.evalsoft.com/DifferentialSurvey/Welcome-Screen.asp?passkey=\[CustomData\]](http://evalsoft07.evalsoft.com/DifferentialSurvey/Welcome-Screen.asp?passkey=[CustomData])

Thank you very much for helping with this important study.

Sincerely,
Winona M. Burt
Doctoral Candidate, Quantitative Methods
Educational Psychology
University of Texas
Austin, TX 78712
Office: 512-385-5520
Mobile: 512-731-5723

Appendix D: Follow-up Email

Dear [FirstName] -

About one week ago I sent an email containing an electronic survey that asked about the meanings of labels used to describe student performance on standardized tests across the United States. To the best of my knowledge, as of Friday, April 2nd you have not yet completed the survey.

I am writing again because your response is important in helping me get accurate results. You are one of a small group of teachers and administrators sampled to represent the opinion of educators across the nation and your response is critical.

An identification number is associated with each survey link so that I can check you off the list when it is completed. At the end of the study the list of names will be destroyed so that individual names can never be connected to the results in anyway.

I hope that you will find the time to complete the questionnaire soon.
Please follow the link below to complete the survey:
[http://evalsoft07.evalsoft.com/DifferentialSurvey/Welcome-Screen.asp?passkey=\[customdata\]](http://evalsoft07.evalsoft.com/DifferentialSurvey/Welcome-Screen.asp?passkey=[customdata])

Please call at the number below if you have any questions at all.
Thank you for your time.

Sincerely,
Winona M. Burt
Doctoral Candidate
Educational Psychology
University of Texas
Austin, TX 78712
Office: 512-385-5520
Mobile: 512-731-5723

Appendix E: Follow-up Postcard

[Date]

Two weeks ago an electronic survey was emailed to you seeking your opinion about the meanings of labels used to describe student performance on standardized tests across the United States. Your name was randomly selected from teachers in your school.

If you have already completed the survey, please accept my sincere thanks. If not, please do so today. I am especially grateful for your help because it is only by asking teachers like you to share your opinions that we can better understand opinions about performance category labels used in high stakes testing.

If you did not receive the first electronic survey, or if it was deleted, please send an email to wburt@mail.utexas.edu or call 512-385-5520 and I will send you the original email. If you prefer, enter the URL below to go directly to the online survey

[http://evalsoft07.evalsoft.com/DifferentialSurvey/Welcome-Screen.asp?passkey=\[customdata\]](http://evalsoft07.evalsoft.com/DifferentialSurvey/Welcome-Screen.asp?passkey=[customdata])

Winona M. Burt
Doctoral Candidate
Educational Psychology
University of Texas
Austin, TX 78712

P.S. If you have any questions at all please do not hesitate to call me at 512-385-5520.

Appendix F: Interview questions

1. Talk to me about your role as [Staff] during the standard setting sessions for [testing program]. What year did you participate? What subject?
2. Let us talk about training. How much of the training was done at the large group level what was done at the small group level?
3. Tell me a about the development of performance category names. Were the terms already established? Were you involved in that process?
4. How were participants selected?
5. What were some of the most common questions?
6. What happens if people just do not seem to get it? What happens to their judgments?
7. Tell me about the communication within groups regarding the performance category names assigned at each level.
8. Let us talk about how the performance category names were referenced during the standard setting session.
9. How often were the definitions referenced?
10. What was reference most often, the performance category names or the definitions?
11. Were [participants/judges] instructed to visualize a student representing each level of performance?
12. Was there any discussion among the group about what the terms meant in their own words or to them?
13. Describe your feelings on the use of labels to describe student performance.

Appendix G: Factor Analysis Correlation Matrix and Standard Deviations

	item1	item2	item3	item4	item5	item6	item7	item8	item9
1: Good-Bad	1.00								
2: Pleasant- Unpleasant	0.228	1.00							
3: Valuable- Worthless	0.249	0.821	1.00						
4: Weak-Strong	-.245	-.719	-.774	1.00					
5: Small-Large	-.163	-.637	-.596	0.584	1.00				
6: Light-Heavy	-.118	-.496	-.499	0.489	0.721	1.00			
7: Sharp-Dull	0.214	0.704	0.714	-.769	-.686	-.564	1.00		
8: Active-Passive	0.050	0.166	0.173	-.166	-.116	-.084	0.299	1.00	
9: Fast-Slow	0.001	0.043	0.095	-.071	-.062	-.128	0.060	0.011	1.00
Standard Deviations:	1.67	1.49	1.47	1.64	1.15	1.05	1.41	1.45	1.39
N = 1170									

References

- Angoff, W. H. (1971). Scales, norms, and equivalent scores. In R. L. Thorndike (Ed.), *Educational Measurement* (2nd ed., pp. 508-600). Washington, D.C.: American Council on Education.
- Asch, S. E. (1946). Forming impressions of personality. *Journal of Abnormal and Social Psychology*, *41*, 258-290.
- Asch, S. E. (1948). The doctrine of suggestion, prestige and imitation in social psychology. *Psychological Review*, *55*, 250-277.
- Bakhtin, M. M. (1984). Problems of Dostoevsky's poetics. In C. Emerson (Ed.). Minneapolis, MN: University of Minnesota Press.
- Bakhtin, M. M. (1986). *Speech genres and other late essays*. Austin, TX: University of Texas Press.
- Beck Evaluation & Testing Associates, Inc., NCS Pearson, in cooperation with Texas Education Agency (2002, July). *Standard setting implementation plan for the Texas assessment of knowledge and skills*. Retrieved November 8, 2003, <http://www.tea.state.tx.us/student.assessment/taks/standards/beta-app-co-pdf>.
- Beretvas, S. N. (2004). Comparison of Bookmark difficulty locations under different item response models. *Applied Psychological Measurement*, *28*(1), 25-47.
- Beretvas, S. N., & Whittaker, T. (2002, April). *Consistency of Bookmark standard setting outcomes: Bookmark difficulty locations and proficiency levels*. Paper presented at the meeting of the American Educational Research Association, New Orleans, LA.
- Berryman-Fink, C., & Verderber, K. S. (1985, March). *Attributions of the term feminist: A factor analytic development of a measuring instrument*. *Psychology of Women Quarterly*, *9*, 51-64.
- Bowler, M. (2003, November 16). *A tale of two test scores*. Sun Spot.net. Retrieved November 19, 2003, <http://www.sunspot.net/news/local/balmd.edbeat16nov16,1,7000923.story>
- Bruner, J. S. (1986). *Actual minds, possible worlds* (pp. 121-133). Cambridge, MA: Harvard University Press.

- Buckendahl, C. W., Smith, R. W., Impara, J. C., & Plake, B. S. (2001, April). *A comparison of Angoff and Bookmark standard setting methods*. Paper presented at the meeting of the National Council on Measurement in Education, Seattle, WA.
- Burke, K. (1965). *Permanence and change: An anatomy of purpose*. Indianapolis: Bobbs-Merrill Company.
- Caron, J. (1992). *An introduction to psycholinguistics*. Toronto, Canada: University of Toronto Press.
- Cizek, G. J. (1993). Reconsidering standards and criteria. *Journal of Educational Measurement, 30*, 93-106.
- Cohen, A. S., Kane, M. T., & Crooks, T. J. (1999). A generalizable examinee-centered method for setting standards on achievement tests. *Applied Measurement in Education, 12*(4), 367-381.
- Crocker, L., & Zieky, M. (1995). *Joint conference on standard setting for large-scale assessments, Executive summary [Vol. 1]*. Aspen Systems Corporation with the National Assessment Governing Board and the National Center for Education Statistics, Washington, D.C.
- Fillmore, C. J. (1971). Types of lexical information. In D. D. Steinberg and L.A. Jakobovits (Eds.), *Semantics: An interdisciplinary reader in philosophy, linguistics, and psychology*. Cambridge: Cambridge University Press.
- Finstuen, K. (1977). Use of Osgood's semantic differential. *Psychological Reports, 41*, 1219-1222.
- Gaskins, R. W. (1996, October/November/December). That's just how it was: The effect of issue-related emotional involvement on reading comprehension. *Reading Research Quarterly, 31*(4), 386-405.
- Gay, W. O. (1971). The bipolar model as it relates to the evaluative factor on the semantic differential. Unpublished doctoral dissertation, University of Texas, Austin.
- Giraud, G., Impara, J. C., & Plake, B.S. (2000, April). *A qualitative examination of teacher's conception of the just competent examinee in Angoff (1971) workshops*. Paper presented at the annual meeting of the American Educational Research Association, New Orleans, LA.
- Green, D. R., Trimble, C. S., & Lewis, D. M. (2003). Interpreting the results of three different standard-setting procedures. *Educational Measurement: Issues and Practice, Spring, 2003*, 22-32.

- Hambleton, R. K. (2001). Setting performance standards on educational assessments and criteria for evaluating the process. In G. J. Cizek (Ed.), *Setting performance standards: Concepts, methods, and perspectives* (pp. 89-117). Mahwah, NJ: Lawrence Erlbaum Associates, Publishers.
- Hertz, N. R., & Chinn, R. N. (2002, April). *The role of deliberation style in standard setting for licensing and certification examinations*. Paper presented at the annual meeting of the National Council on Measurement in Education, New Orleans, LA.
- Hoel, T. L. (1997). Voices from the classroom. *Teaching and Teacher Education*, 13(1), 5-16.
- Hu, L., & Bentler, P.M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling*, 6(1), 1-55.
- Impara, J. C., & Plake, B.S. (1977). Standard setting: An alternative approach. *Journal of Educational Measurement*, 34(4), 353-366.
- Improving America's School: *A Newsletter on Issues in School Reform*, Spring, 1-8. Retrieved June 17, 2003, <http://www.ed.gov/pubs/IASA/newsletters/standards/pt2.html>
- Jacobson, M. B. (1979). A rose by any other name: Attitudes toward feminism as a function of its label. *Sex Roles*, 5(3), 365-371.
- Kane, M. T. (2001). So much remains the same: Conception and status of validation in setting standards. In G. J. Cizek (Ed.), *Setting performance standards: Concepts, methods, and perspectives* (pp. 53-88). Mahwah, NJ: Lawrence Erlbaum Associates, Publishers.
- Kerlinger, F.N. (1986). *Foundations of behavioral research* (3rd. ed). New York: Holt, Rinehart & Winston.
- Kintsch, W. (1994, April). Text comprehension, memory, and learning. *American Psychologist*, 49(4), 294-303.
- Kintsch, W. & Van Dijk, T. A. (1978, September). Toward a model of text comprehension and production. *Psychological Review*, 85(5), 363-394.
- Kiplinger, V.L. (1997). Standard setting procedures for the specification of performance levels on a standards-based assessment. Retrieved on August 21, 2003, <http://www.cde.state.co.us/cdeassess/csap/asperf.htm>
- Kress, G. (1989). *Linguistic processes in sociocultural practice*. Oxford, England: Oxford University Press.

- Lewis, D. M., & Mitzel, H.C. (1995, September). An item response theory based standard setting procedure. In D. R. Green (Chair), Some uses of item response theory in standard setting. Symposium conducted at the annual meeting of the California Educational Research Association, Lake Tahoe, NV.
- Lewis, D. M., Mitzel, H. C., Green, D. R., (1996, June). *Standard setting: A bookmark approach*. Paper presented at the Council of Chief State School Offices National Conference on Large-Scale Assessment, Boulder, CO.
- Lewis, D. M., Green, D. R., Mitzel, H. C., Baum, K. & Patz, R. J. (1999). *The Bookmark standard setting procedure: Methodology and recent implementations*. Manuscript submitted for publication.
- Linn, R. L. (1994, October). *The likely impact of performance standards as a function of uses: From rhetoric to sanctions*. Paper presented at the National Center for Education Statistics and National Assessment Governing Board Joint Conference on Standard Setting for Large-Scale Assessments,
- Livingston, S. A., & Ziemy, M. J. (1982). Passing scores: A manual for setting standards of performance on educational and occupational tests.
- Loftus, E. F. (1973). Activation of semantic memory. *American Journal of Psychology*, 86, 331-337.
- Loftus, E. F., & Palmer, J. C. (1974). Reconstruction of automobile destruction: An example of the interaction between language and memory. *Journal of Verbal Learning and Verbal Behavior*, 13, 585-589.
- Loomis, S.C., & Bourque, M.L. (2001). From tradition to innovation: Standard setting on the national assessment of educational progress. In G.J. Cizek (Ed.), *Setting performance standards: Concepts, methods, and perspectives* pp. (175-217). Mahway, NJ: Erlbaum.
- Mehrens, W. A. (1995). Methodological issues in standard setting for educational exams. In *Proceedings of Joint Conference on Standard Setting for Large-Scale Assessments* (pp. 221-263).
- Murphy, S. T., & Zajonc, R. B. (1993). Affect, cognition, and awareness: Affective priming with optimal and suboptimal stimulus exposures. *Journal of Personality and Social Psychology*, 64, 723-739.
- National Assessment Governing Board. (2002, March). Using the national assessment of educational progress to confirm state test results. Retrieved December 9, 2003, from http://www.nagb.org/pubs/color_document.pdf
- Nunnally, J. C. (1967). *Psychometric theory* (pp. 514-550). New York: McGraw-Hill.

- Olson, L. (2003) Approved is relative term for education department. *Education Week*. Retrieved September 4, 2003 from <http://www.educatonweek.org/ew/ewstory.cfm?slug=43account.h22&keywords=approved>
- Osgood, C. E., Suci, G. J., & Tannenbaum, P. H. (1957). *The measurement of meaning*. Urbana, IL: University of Illinois.
- Perrig, W. & Kintsch, W. (1985). Propositional and situational representations of text. *Journal of Memory and Language*, 24, 503-518.
- Reckase, M. D. (1994, June). *Standard setting on performance assessments: A comparison between the paper selection method and the contrasting groups method*. Paper presented at the National Conference on Large Scale Assessment, Albuquerque, NM.
- Shepard, L., Glaser, R., Linn, R., & Bohrnstedt, G. (1993). *Setting performance standards for student achievement*. Stanford, CA: National Academy of Education.
- Stevens, J. P. (2002). Applied multivariate statistics for the social sciences (4th ed.), Multivariate planned comparisons on SPSS MANOVA (pp. 231-248). Mahwah, NJ: Erlbaum.
- Taylor, J. R. (2002). Near synonyms as co-extensive categories: 'High' and 'tall' revisited. *Language Sciences*, 25, 263-284.
- U.S. Department of Education. (2001). *Executive summary of the No Child Left Behind Act of 2001*. Retrieved September 10, 2003, from <http://www.ed.gov/print/nclb/overview/intro/execsumm.html>
- U.S. Department of Education. (2002). Excerpts from consolidated state application. Retrieved July 27, 2004, from <http://www.educationadvisor.com/documents/>
- U.S. Department of Education. (2004). Approved state accountability plans. Retrieved April 01, 2004, from <http://www.ed.gov/admins/lead/account/stateplans03/index.html>
- Wilson, T. D., Dunn, D. S., Kraft, D. & Lisle, D. J. (1989). Introspection, attitude change, and attitude-behavior consistency: The disruptive effects of explaining why we feel the way we do. In L. Berkowitz (Ed.), *Advances in experimental social psychology* (Vol. 22, pp. 287-343). New York: Academic Press.
- Wurm, L. H., & Vakoch, D. A. (2000). The adaptive value of lexical connotation in speech perception. *Cognition and Emotion*, 14(2), 177-191.

Zieky, M. J. (2001). So much has changed: How the setting of cutscores has evolved since the 1980s. In G. J. Cizek (Ed.), *Setting performance standards: Concepts, methods, and perspectives* (pp.19-52). Mahwah, NJ: Erlbaum.

Vita

Winona Madelain Burt was born in Baytown, Texas on August 12, 1977, the daughter of Winston Burt and Gweneth Burt. After completing her work at Robert E. Lee High School, Baytown, Texas, in 1995, she entered St. Edward's University in Austin, Texas. She received her Bachelor of Arts in Psychology from St. Edward's University in 1999 and began working with Evaluation Software Publishing (ESP) Incorporated the same year. She worked with ESP for 5 years in the research and evaluation department. She began taking courses for the Doctorate in the fall of 1999. She received a Doctorate in Educational Psychology with a major in Quantitative Methods from The University of Texas, Austin in August 2004 and is moving to Washington, D.C. the summer of 2004 to begin work as a research scientist at the American Institutes for Research (AIR).

Permanent Address: 303 Welford Lane, Highlands, TX 77562

This dissertation was typed by the author.