

2002

Investigation of two standard setting methods for a licensure examination.

Mary J. Pitoniak

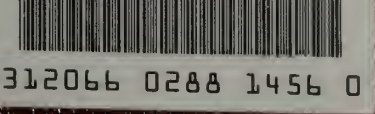
University of Massachusetts Amherst

Follow this and additional works at: <https://scholarworks.umass.edu/theses>

Pitoniak, Mary J., "Investigation of two standard setting methods for a licensure examination." (2002). *Masters Theses 1911 - February 2014*. 2386.

Retrieved from <https://scholarworks.umass.edu/theses/2386>

This thesis is brought to you for free and open access by ScholarWorks@UMass Amherst. It has been accepted for inclusion in Masters Theses 1911 - February 2014 by an authorized administrator of ScholarWorks@UMass Amherst. For more information, please contact scholarworks@library.umass.edu.



312066 0288 1456 0

INVESTIGATION OF TWO STANDARD SETTING METHODS
FOR A LICENSURE EXAMINATION

A Thesis Presented

by

MARY J. PITONIAK

Submitted to the Graduate School of the
University of Massachusetts Amherst in partial fulfillment
of the requirements for the degree of

MASTER OF SCIENCE

February 2002

Department of Psychology

© Copyright by Mary J. Pitoniak 2002

All Rights Reserved

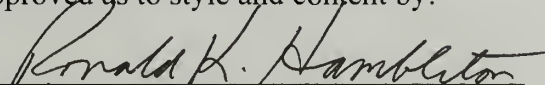
INVESTIGATION OF TWO STANDARD SETTING METHODS
FOR A LICENSURE EXAMINATION

A Thesis Presented

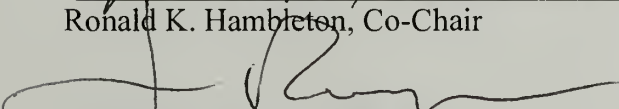
by

MARY J. PITONIAK

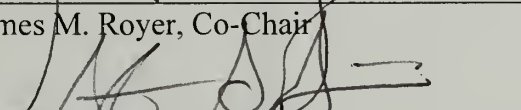
Approved as to style and content by:



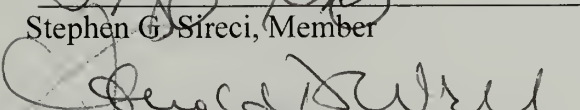
Ronald K. Hambleton, Co-Chair




James M. Royer, Co-Chair



Stephen G. Sireci, Member



Arnold D. Well, Member



Melinda Novak, Chair
Department of Psychology

ACKNOWLEDGEMENTS

I have had the fortune of benefiting from the wisdom and support of faculty from two distinguished departments at the University of Massachusetts Amherst: the Research and Evaluation Methods Program and the Department of Psychology. I would like to thank the members of my committee, each of whom has facilitated the completion of my coursework and this master's thesis.

First, I thank Ron Hambleton, who since the first course I took as a non-degree student four years ago has inspired me with both his command of psychometric knowledge and his love of the field. I have greatly appreciated his mentorship and true concern for my progress in the program. Ron's ability to be closely involved with the activities of his students despite his stature as a national expert is impressive. Next I thank Steve Sireci, who provided me with my initial opportunity to learn about standard setting, an area in which my interest has only grown since that time. Steve's tireless enthusiasm and gracious assistance have been a source of great support to me.

Third, I extend my appreciation to Mike Royer, whose willingness to sponsor my entry into the Psychology Department revived a long-standing dream that my graduate education would be tied to this field. Although the main focus of my interest is now numbers and test-takers, and not psychiatric patients, I am grateful that my psychology background is now serving as a foundation for my current work. I have learned a great deal from Mike and appreciate his working with me on my first in-depth literature review after returning to school, which most certainly would not have been published without his support and advice. Fourth, I thank Arnie Well, who provided me

with a solid refreshing of rusty statistical concepts in a kind and skilled manner that not only lessened my anxiety but strengthened my confidence in my abilities.

I owe a large debt of gratitude to the American Institute of Certified Public Accountants (AICPA) for supporting the research on which this thesis is based. I thank Craig Mills, Jerry Melican, Bruce Biskin, Josiah Evans, Ahava Goldman, and Adell Battle for their invaluable assistance in providing funding, locating participants, finalizing materials, and hosting the panel meetings.

I would like to acknowledge as well the financial support and professional opportunities provided to me by National Evaluation Systems, Inc. (NES), particularly in the early years of my graduate studies. My work at NES laid the foundation for my interest in and pursuit of a graduate degree in psychometrics, for which I will always be grateful, and I count as valued friends many of the staff there.

Fellow graduate students are invariably an invaluable force in continuing one's progress through school, and my case is no different. Special thanks go to Mike Jodoin, who also served as a co-facilitator for the study. I also wish to thank Lisa Keller, April Zenisky, Dean Goodman, and Billy Skorupski for their ability to make me laugh and persevere in the face of challenges. I thank as well my friends Bob and Max, both of whom encouraged me when the going got rough to remain dogged in pursuit of my goals.

Lastly, I thank my parents for their providing me with the resources—genetic, emotional, and financial—needed to be successful in both graduate school and in life in general. I greatly appreciate their support and faith in my abilities, and I dedicate this thesis to them.

ABSTRACT

INVESTIGATION OF TWO STANDARD SETTING METHODS FOR A LICENSURE EXAMINATION

FEBRUARY 2002

MARY J. PITONIAK, B.A., SMITH COLLEGE

M.S., UNIVERSITY OF MASSACHUSETTS AMHERST

Directed by: Professors Ronald K. Hambleton and James M. Royer

Rather than being merely the last step in the development of a professional licensure exam, standard setting should provide critical validity evidence that examination scores are appropriate for the purpose of awarding a license. However, standard setting is often regarded as the weakest link in the chain of validity evidence. Contemporary licensure and certification tests typically include multiple-choice items and performance tasks, which causes some problems when implementing traditional standard setting methods. Furthermore, standard setting panelists in many professional areas have hectic schedules and are both expensive and difficult to recruit. Therefore, standard setting methods are needed that are psychometrically defensible, but minimize the amount of time needed from expert panelists.

In this study two relatively new standard setting methods designed for today's complex assessments were implemented: the Item Cluster method and the Direct Consensus method. Each of these methods was used previously with large-scale credentialing exams (with promising results in both cases), but this study represents the first comparison between the methods.

Data obtained in the study were evaluated within Kane's (1994, 2001) validity framework, in which three sources of evidence are considered: procedural, internal, and external. Major findings related to consistency within a method and across methods. The Direct Consensus method yielded inconsistent cut scores across sessions, while the Item Cluster method produced consistent cut scores. Comparisons across the two methods revealed cut scores that were quite different from each other. The Direct Consensus method yielded higher cut scores, which resulted in estimated passing rates that were more in line with operational trends than those of the Item Cluster method. In general, panelists felt more positively about the Direct Consensus method; in addition, that method takes substantially less time to implement.

Both methods appear promising, but future research should focus on those aspects of each method that provoked the most concern. For the Direct Consensus method, inconsistency of cut scores should be the focus. For the Item Cluster method, the minimal degree to which panelists said they used complete examinee profile information (a key component of the method) and the low cut scores set should be investigated further.

TABLE OF CONTENTS

	Page
ACKNOWLEDGEMENTS	iv
ABSTRACT	vi
LIST OF TABLES	xi
LIST OF FIGURES	xiv
CHAPTER	
1. INTRODUCTION.....	1
1.1 Background.....	1
1.2 Standard Setting	4
1.3 Statement of the Problem.....	7
1.4 Purpose of the Study.....	9
1.5 Significance of the Problem.....	12
2. REVIEW OF LITERATURE	14
2.1 Steps in Standard Setting	14
2.2 Classification Dimensions	18
2.3 Traditional Standing-Setting Methods	21
2.3.1 Angoff Method.....	21
2.3.2 Ebel Method.....	24
2.3.3 Nedelsky Method	25
2.3.4 Jaeger Method	26
2.3.5 Borderline Group Method.....	27
2.3.6 Contrasting Groups Method.....	27
2.4 Newer Standing-Setting Methods	29
2.4.1 Cluster Analysis Method	29
2.4.2 Judgmental Policy Capturing Method	31
2.4.3 Dominant Profile Method	32
2.4.4 Bookmark Method	33
2.4.5 Extended Angoff Method	35
2.4.6 Analytic Judgment Method.....	37
2.4.7 Body of Work Method.....	38
2.4.8 Direct Consensus Method.....	39
2.4.9 Item Cluster Method.....	41

2.5	Evaluation Criteria	45
2.6	Summary.....	46
3.	METHODOLOGY	49
3.1	Study Design.....	49
3.2	Panelists.....	49
3.3	Test Items.....	51
3.4	Meeting Procedures.....	53
	3.4.1 Orientation and Training.....	53
	3.4.2 Direct Consensus Method.....	55
	3.4.3 Item Cluster Method.....	57
	3.4.4 Collect Evaluation Information.....	60
4.	RESULTS	61
4.1	Direct Consensus Method.....	61
	4.1.1 Detailed Panelist Rating Information	61
4.2	Item Cluster Method	71
	4.2.1 Boundary Method.....	71
	4.2.2 Regression Method.....	74
	4.2.3 Equating Method.....	79
	4.2.4 Detailed Panelist Rating Information	80
4.3	Summary Information	106
	4.3.1 Comparison of Cut Scores and Their Impact.....	106
	4.3.2 Timing Information	110
	4.3.2 Evaluation Survey Results	113
5.	DISCUSSION	141
5.1	Evaluation of Results Within Validity Framework.....	141
	5.1.1 Procedural Evidence of Validity	141
	5.1.2 Internal Evidence of Validity.....	149
	5.1.3 External Evidence of Validity.....	155
5.2	Conclusions.....	158
5.3	Future Research.....	160

APPENDICES

A. SAMPLE ITEM RATING FORMS 161
B. EVALUATION SURVEY 164
REFERENCES 173

LIST OF TABLES

Table	Page
2.1 Hambleton’s (1998) Steps for Standard Setting	15
2.2 Direct Consensus and Item Cluster Methods: Similarities And Differences	44
2.3 Summary of Criteria for Evaluating Standard-Setting Procedures	47
3.1 Experimental Design	49
3.2 Characteristics of Panel Members	51
3.3 Meeting Agenda	54
3.4 Sample Panelist Data Display For the Direct Consensus Method	56
3.5 Distribution of Candidate Response Strings for Item Cluster Method.....	58
3.6 Rating Scale for Item Cluster Method.....	59
4.1 Panelist Cut Scores Across Rounds for Direct Consensus Method: Panel B (Morning).....	63
4.2 Panelist Cut Scores Across Rounds for Direct Consensus Method: Panel A (Afternoon)	64
4.3 Descriptive Statistics for Direct Consensus Method Ratings: Panel B (Morning).....	65
4.4 Descriptive Statistics for Direct Consensus Method Ratings: Panel A (Afternoon)	66
4.5 Relationship Between Mean Cluster <i>p</i> -value and Panelist Ratings for Direct Consensus Method: Panel B (Morning).....	67
4.6 Relationship Between Mean Cluster <i>p</i> -value and Panelist Ratings for Direct Consensus Method: Panel A (Afternoon).....	68
4.7 Cut Scores for Panel A (Morning): Item Cluster Method	72
4.8 Cut Scores for Panel B (Afternoon): Item Cluster Method.....	73

4.9	Percentage of Variance Accounted for by Regression Model: Panel A (Morning).....	76
4.10	Percentage of Variance Accounted for by Regression Model: Panel B (Afternoon)	76
4.11	Difference in Cut Score and Percentage of Variance Accounted for by Regression Model When Using Rescaled Ratings: Panel A (Morning).....	77
4.12	Difference in Cut Score and Percentage of Variance Accounted for by Regression Model When Using Rescaled Ratings: Panel B (Afternoon)	78
4.13	Descriptive Statistics for Item Cluster Method Ratings: Panel A (Morning), Cluster 1	82
4.14	Descriptive Statistics for Item Cluster Method Ratings: Panel A (Morning), Cluster 2	83
4.15	Descriptive Statistics for Item Cluster Method Ratings: Panel A (Morning), Cluster 3	84
4.16	Descriptive Statistics for Item Cluster Method Ratings: Panel A (Morning), Cluster 4	85
4.17	Descriptive Statistics for Item Cluster Method Ratings: Panel A (Morning), Cluster 5	86
4.18	Descriptive Statistics for Item Cluster Method Ratings: Panel B (Afternoon), Cluster 1	87
4.19	Descriptive Statistics for Item Cluster Method Ratings: Panel B (Afternoon), Cluster 2	88
4.20	Descriptive Statistics for Item Cluster Method Ratings: Panel B (Afternoon), Cluster 3	89
4.21	Descriptive Statistics for Item Cluster Method Ratings: Panel B (Afternoon), Cluster 4	90
4.22	Descriptive Statistics for Item Cluster Method Ratings: Panel B (Afternoon), Cluster 5	91
4.23	Changes in Ratings by Panelist Across Rounds for Item Cluster Method: Panel A (Morning)	92
4.24	Changes in Ratings by Panelist Across Rounds for Item Cluster Method: Panel B (Afternoon)	93

4.25	Correlations Between Examinee Profile Scores and Panelist Ratings for Item Cluster Method: Panel A (Morning).....	94
4.26	Correlations Between Examinee Profile Scores and Panelist Ratings for Item Cluster Method: Panel B (Afternoon)	95
4.27	Comparison of Cut Scores Across Methods.....	109
4.28	Summary of Timing Information	110
4.29	Timing Information: Direct Consensus Method.....	111
4.30	Timing Information: Item Cluster Method.....	112
4.31	Evaluation Survey Results: General Questions.....	114
4.32	Evaluation Survey Results: Direct Consensus Method	116
4.33	Evaluation Survey Results: Item Cluster Method	120
4.34	Evaluation Survey Results: Open-Ended Questions	125

LIST OF FIGURES

Figure	Page
4.1	Mean panelist cluster rating, in terms of percentage of items, and mean cluster <i>p</i> -value for Panel B (Morning), Round 2, Direct Consensus method. 69
4.2	Mean panelist cluster rating, in terms of percentage of items, and mean cluster <i>p</i> -value for Panel A (Afternoon), Round 2, Direct Consensus method. 70
4.3	Panelist ratings and examinee profile scores for Panel A (Morning), Cluster 1, Round 2, Item Cluster method. 96
4.4	Panelist ratings and examinee profile scores for Panel A (Morning), Cluster 2, Round 2, Item Cluster method. 97
4.5	Panelist ratings and examinee profile scores for Panel A (Morning), Cluster 3, Round 2, Item Cluster method. 98
4.6	Panelist ratings and examinee profile scores for Panel A (Morning), Cluster 4, Round 2, Item Cluster method. 99
4.7	Panelist ratings and examinee profile scores for Panel A (Morning), Cluster 5, Round 2, Item Cluster method. 100
4.8	Panelist ratings and examinee profile scores for Panel B (Afternoon), Cluster 1, Round 2, Item Cluster method. 101
4.9	Panelist ratings and examinee profile scores for Panel B (Afternoon), Cluster 2, Round 2, Item Cluster method. 102
4.10	Panelist ratings and examinee profile scores for Panel B (Afternoon), Cluster 3, Round 2, Item Cluster method. 103
4.11	Panelist ratings and examinee profile scores for Panel B (Afternoon), Cluster 4, Round 2, Item Cluster method. 104
4.12	Panelist ratings and examinee profile scores for Panel B (Afternoon), Cluster 5, Round 2, Item Cluster method. 105

CHAPTER 1

INTRODUCTION

1.1 Background

Educational tests are often employed so that a decision can be made about an individual's level of competence in a given domain. Standard setting, which involves the establishment of a cut score that discriminates between levels of performance so that these decisions can be made, is thus a significant part of the overall assessment development and implementation process. Rather than being a minor last step in this process, appropriate and effective standard-setting activities provide a critical link in the chain of validity evidence to be gathered in order to make sound interpretations from test scores.

However, standard setting is not an objective psychometric process in which a study is conducted in order to estimate the value of a true population parameter (Cizek, 1996a). As Hambleton (1998) observed, "it is well known that there are no true performance standards waiting to be discovered through research studies. Rather, setting performance standards is ultimately a judgmental process" (p. 87). As a result, standard setting has been the subject of debate as psychometricians and policy makers have struggled with the best way to conduct an activity that has such subjective features. While some have decried the "arbitrary" nature of the activity and outcome (Glass, 1978), others have stressed that the process itself can be made procedurally sound, thus providing evidence that the standard itself is defensible and credible (e.g., Cizek, 1993; Hambleton, 1998; Kane, 1994, 2001).

Several factors have led to increased attention being given to establishing the credibility of existing standard-setting methods and developing new methods. The first factor is the degree to which high-stakes testing has become more prevalent in recent years (Cizek, 2001). National attention has been given to the impact of statewide testing programs whose effects include determining the graduation status of children, as well as whether schools will be subjected to state takeover or their teachers will get bonuses (Morse, 2000). Although at least one survey has shown that such tests are supported by the public (Business Roundtable, 2000), criticisms still arise over the soundness and use of large-scale assessments. As high-stakes tests are increasingly used to make both educational decisions and determinations of status in other areas such as licensure, the importance of establishing the soundness of standard-setting methods that establish cut scores used for these decisions is paramount (American Educational Research Association, 2000).

Two additional factors have been noted by Berk (1996) as contributing to interest in refining standard-setting techniques. One factor is closely linked to the increased use of statewide educational testing programs described above. Such testing programs often have multiple cut scores (for example, the Massachusetts state-wide testing program requires three cut scores to distinguish among the four performance levels of Advanced, Proficient, Needs Improvement, and Failing; Massachusetts Department of Education, 2000). The use of multiple cut scores on a test may magnify any inherent problems in a particular standard-setting method, including how to define the characteristics of the candidates above and below each passing standard.

Another factor cited by Berk (1996) as stimulating renewed interest in standard setting is the increasing use of polytomous item formats. Many standard-setting methods were developed for use with multiple-choice questions; however, since many assessments implemented in recent years also utilize constructed-response items or performance assessments, those old methods may not suffice (see also Hambleton, Jaeger, Plake, & Mills, 2000).

As a result of these factors, standard-setting research has proliferated in the last decade. Some of these new methods have been implemented operationally, while some are still in the research phase. The proliferation of new methods is very positive. However, there is a need to continue to subject all new approaches to empirical research in order to determine their strengths and weaknesses.

In addition, it should be noted that the multiple-choice format is still the item type of choice for many credentialing examinations (Meara, 2000). Hence, continued development and scrutiny of methods that may be used for tests comprised in whole or in part of multiple-choice items is warranted. Since the formulation of new methods for use with multiple-choice items has been rather limited in the past few decades (with the few exceptions being the Jaeger [1982] method and the bookmark method [Lewis, Mitzel, & Green, 1996], both of which will be reviewed in Chapter 2), it is important to keep in mind the need to develop and refine methods suited for that item type. Therefore, implementation in a controlled study of two new methods, both of which can be used for tests with multiple-choice items as well as selected-response items, is a useful extension of recent research.

1.2 Standard Setting

In both educational settings and other areas in which high-stakes tests are administered, a distinction can be made between two types of standards: performance standards and passing standards. Performance standards are descriptions of the desired level of proficiency to be represented by scores within a given range. As such, a performance standard represents “the required level of achievement specified in terms of what candidates need to know and be able to do . . . [it] is a qualitative description of the level of achievement on the [knowledge, skills, and abilities] needed for practice at a particular level” (Kane, Crooks, & Cohen, 1997, p. 5).

In contrast, passing standards are specific cut scores on the score scale that serve to separate individuals into the categories described by the performance standards (Kane et al., 1997). Passing standards may distinguish between two groups (as in a pass/fail or mastery/nonmastery distinction) or among three or more groups (as with assignment of proficiency levels). Kane (1994, 2001) noted that the establishment of performance standards, the first part of the overall process, involves policy decisions. In contrast, operationalizing those performance standards as passing standards, the second part of the overall process, is the province of the standard-setting study.

Standard-setting methods vary along several dimensions, the most basic of which is the distinction between test-centered methods and examinee-centered methods (Cizek, 1996a; Jaeger, 1989; Kane, 1994). Test-centered standard-setting methods involve the formation of judgments about test content. The approach employs standard-setting panels to carefully review test items and provide judgments regarding expected levels of performance on each item by specific types of test takers. The notion

of a borderline examinee is a fundamental component of test-centered standard-setting methods¹. The borderline examinee is someone who possesses “just enough” knowledge, skill, or ability to meet a particular performance standard. Test-centered standard-setting methods require panelists to provide ratings regarding how well borderline examinees are likely to do on each item. For tests comprising multiple standards, a different borderline examinee must be envisioned for each performance standard.

In contrast to test-centered standard-setting methods, examinee-centered standard-setting methods focus on examinees rather than on test items. Standard-setting panelists are used to classify examinees into performance categories, such as “pass,” “fail,” or “borderline.” The panelists used to categorize examinees depend on the method chosen. One option is to have standard-setting panelists identify borderline examinees on the basis of their knowledge of the examinees, after which the test scores for these borderline examinees are gathered and their median test score is typically used as the cut score (borderline group method). Another is to have panelists identify two different groups of examinees, one whose members are clearly above a particular standard and another whose members are clearly below that standard. The test score distributions of these two groups are then contrasted to select the cut score (contrasting group method).

¹In this study the term “borderline examinee” will generally be used. However, several other terms may be found in the literature to describe this type of person, depending in part on the assessment context. These other terms include “minimally acceptable person” (Angoff, 1984), “minimally competent candidate” (Plake, Impara, & Irwin, 1999); “borderline student” (Hambleton, 1998b); “borderline test-taker” (Livingston & Zieky, 1982); and “just barely certifiable candidate” (Hambleton & Plake, 1995).

New standard-setting methods being developed can be classified in terms of these two dimensions (test-centered vs. examinee-centered) as well as by other features of the process to be described in Section 2.2 (e.g., Hambleton, Jaeger, Plake, & Mills, in press). However, regardless of the specific methods employed in standard-setting studies, there are many similarities in the steps that are implemented. Hambleton (1998) has outlined 11 typical steps that are employed in a panel-based standard-setting study. They are: (1) choose a panel that is large and representative of the stakeholders; (2) choose a standard-setting method, prepare training materials, and finalize the meeting agenda; (3) prepare descriptions of the performance categories; (4) train the panelists to use the method, including providing practice in making ratings; (5) compile item ratings or other data from the panelists; (6) conduct a panel discussion, consider actual performance data, and provide feedback on inter-panelist and intra-panelist consistency; (7) compile item ratings a second time, which may be followed by more discussion and feedback; (8) compile panelist ratings and average to obtain the passing standard; (9) present consequences data to the panel; (10) revise, if necessary, and finalize the passing standard(s), and conduct a panelist evaluation of the process itself and their level of confidence in the resulting standard(s); and (11) compile technical documentation to support the validity of the passing standard(s).

The eleventh step listed by Hambleton (1998) is a crucial one. The evidence needed to support valid interpretations of classifications resulting from test score use must be gathered and documented. As noted earlier, there are no absolute criteria against which standards can be validated. Similarly, there are no absolute criteria against which different standard-setting studies may be evaluated (Kane, 1994).

However, since the absence of perfect criteria does not excuse a testing agency from providing evidence that the standards are reasonable and appropriate, several sets of guidelines and recommendations for carrying out a standard-setting study have been formulated (Cizek, 1993, 1996a, 1996b; Hambleton, 1998; Hambleton & Powell, 1983; Jaeger, 1991; Livingston & Zieky, 1982; Norcini & Shea, 1997; Plake, 1997). In addition, the Standards for Educational and Psychological Testing (American Educational Research Association, National Council on Measurement in Education, & American Psychological Association, 1999) stipulate several recommendations for conducting and evaluating standard-setting studies. These criteria and recommendations are important considerations to keep in mind during and after the implementation of a standard-setting study.

1.3 Statement of the Problem

Changes in testing practice necessitate the development of new standard-setting methods. Research is needed into the soundness of new methods, including whether resulting passing standards are replicable across panels. The current study evaluated the standards obtained from two new standard-setting methods—the Direct Consensus method and the Item Cluster method—in terms of established validity criteria. The data from this study were evaluated within the framework proposed by Kane (1994, 2001), in which three general sources of validity evidence are viewed as important: procedural, internal, and external (see Table 2.3, Chapter 2). Using this framework, the following hypotheses were investigated.

1. Procedural evidence. According to Kane (1994), “procedural evidence focuses on the appropriateness of the procedures used and the quality of the implementation of these procedures” (p. 437). Implementation of each standard-setting method was evaluated within the context of five criteria: explicitness, practicability, implementation of procedures, panelist feedback, and documentation. A particularly important source of information about the practicability of standard-setting processes is the panelists themselves (Geisinger, 1991; Kane, 1994, 2001); therefore, feedback was obtained from panelists about how clear they felt the training was and how comfortable they felt with both standard-setting procedures as implemented in this study.
2. Internal evidence. The Standards (AERA et al., 1999) indicate that “whenever feasible, an estimate should be provided of the amount of variation in cut scores that might be expected if the standard-setting procedure were replicated” (p. 60). In this study, each of the two standard-setting methods was utilized by two different panels. The within-session replication afforded by the current study’s design thus allowed for a more direct estimation of the standard error than studies in which only one panel uses a given method (Kane, 1994, 2001). Additional internal evidence related to intrapanelist and interpanelist consistency was also evaluated. Comparisons of both types of consistency across different standard-setting methods are useful (Berk, 1996; Cizek, 1996b).
3. External evidence. Comparisons across methods are also a valuable source of validity evidence (Kane, 1994). Two different standard-setting methods were used in this study, with the same sets of items. This afforded the opportunity to

determine whether the Item Cluster method and Direct Consensus method yielded similar cut scores. In addition, the reasonableness of these cut scores, as reflected in estimated pass rates, was also examined.

1.4 Purpose of the Study

In this study, two new standard-setting methods were investigated in the context of a licensure examination: (1) the Direct Consensus method, and (2) the Item Cluster method. Each of the methods has been used only once before; as such, the current study serves as a much-needed replication of earlier studies. As Norcini (1994) noted, “it is crucial that any such work be set in the context of earlier studies, and replication is highly desirable. Where possible, experimental designs will produce more useful results” (p. 172).

In one of the two approaches used in this study, the Direct Consensus method, panelists set passing standards directly based on a consideration of information that includes the following: descriptions of the performance standards, previous exams and the corresponding standards; the content of the current exam and its scoring rubrics; any statistical data that may be available; and sample examinee constructed responses, if applicable. The facilitator engages panelists in a discussion of all of the available information and attempts to help panelists reach consensus on the resulting passing standards.

The method is termed direct because panelists work with the actual exam scale. It is described as reflecting consensus because the goal is to have the panel arrive at passing standards that they can agree upon (though as a last resort, the mean of their

recommended passing standards can be used). An advantage of this method is that it is faster than other methods because item level ratings are not provided and panelists do not need to sort through rather large sets of student papers. Greater efficiency in the process has been suggested as a research goal for standard-setting by Norcini (1994). This is of particular importance in technology certification areas since tests are updated often due to quickly changing content in those fields (Sireci, Hambleton, Huff, & Jodoin, 2000).

The Direct Consensus method was recently implemented by Sireci, Hambleton, et al. (2000) in a standard-setting study within a certification context. In this study, two different panels used both the Direct Consensus method and the Angoff (1971) method. The two panels' passing scores were more consistent with the former method than with the latter. Also of note is the fact that the Direct Consensus method took slightly less time to implement than the Angoff method. All panelists viewed positively one significant feature of the Direct Consensus method—that panelists have direct control over the final standard (in contrast to other methods, where panelists may not be aware of the final standard and/or be able to adjust their ratings once the group standard is known). The current study provided an opportunity to apply the method in a generally similar setting, but with a different examination, and, of course, with different panelists. Comparison of the Direct Consensus method to another approach other than Angoff's provided additional useful information about this new procedure.

The second standard-setting approach used, the Item Cluster method, also involves dividing an exam into clusters of items. In this method, however, panelists are also presented with patterns of examinee responses to the questions in each cluster.

When an examinee answers a multiple-choice question incorrectly, the panelists are informed of the distractor chosen by the examinee. For any constructed-response questions, panelists see actual student work. Panelists assign these student response patterns to one of six categories, ranging from 1 (hopeless) to 6 (exceptional). After completing their initial ratings, panelists meet to discuss their ratings and then have another opportunity to classify the student response patterns. Arriving at a final set of standards can be handled in one of three ways: (1) by looking at the mean scores of student response patterns assigned by panelists to the borderline categories (boundary method); (2) by fitting a linear or non-linear regression line to the mean scores of examinee response patterns assigned to each of the six performance categories; or (3) by using an “equating” method that entails looking at the relationship between the scores obtained by different examinees and the ratings assigned by panelists in terms of the percentage of both distributions found below specific points.

The Item Cluster method was first implemented in a study comparing it to the Angoff method (Mills, Hambleton, Biskin, Kobrin, Evans, & Pfeffer, 2000). In that study, the test for which standards were set (not operationally, but as part of a research effort) was the Uniform CPA Examination. This method has several advantages that warrant its further investigation: (a) the method can handle both multiple-choice and performance tasks in the same test, (b) it allows panelists to consider the actual performance of students; however, the chunks are small enough that the patterns of rights and wrongs, and actual work on the constructed response questions, can be meaningfully judged holistically, and (c) the method is focused on actual student work—something that panelists often say they want to consider in setting standards.

Results from the Mills et al. (2000) study indicated that the cut scores yielded by the Item Cluster method were more consistent across panels than those obtained with the Angoff method (though this may have been confounded with a facilitator effect). The cut scores, while lower than those set using the Angoff method, were more consistent with the cut scores that resulted when the Beuk (1984) compromise method was utilized. Panelists also felt more positively about the Item Cluster method than about the Angoff method.

In the current study, the results obtained with both the Direct Consensus and Item Cluster approaches were analyzed in terms of the validity criteria outlined by Kane (1994, 2001) and others. The careful analysis of standard-setting methods in terms of these criteria is a critical step in the thorough exploration of new (and existing) procedures.

1.5 Significance of the Problem

The 1999 Standards for Educational and Psychological Testing (AERA et al.) note that the establishment of a cut-point to divide the score scale into categories is a “critical step” in the test development and implementation process (p. 53). As a result, the Standards recommend that “where the results of the standard-setting process have highly significant consequences, and especially where large numbers of examinees are involved, those responsible for establishing cut scores should be concerned that the process by which cut scores are determined be clearly documented and defensible” (p. 54).

As noted earlier, the increased prevalence of high-stakes testing as a component of the decision process in education has certainly led to the presence of “highly significant consequences” for these tests (Cizek, 2001). Similarly, tests for licensure and certification are of very high consequence for potential practitioners whose career may hinge on a test score.

Since many of these high-consequence tests are comprised of newer assessment formats in addition to multiple-choice questions, and also have more than one cut score, research into new standard-setting methods is essential. The validity of interpretations made from test scores rests in part on the credibility of the standard-setting methods used. A sound research base is an important step in the establishment of that credibility, and studies such as the current one contribute to that crucial foundation.

CHAPTER 2

REVIEW OF LITERATURE

Within this section, steps in the standard-setting process and dimensions along which standard-setting methods can be classified are described. Then, both older methods and newer methods are outlined. In addition, criteria by which the validity of passing standards may be assessed are presented.

2.1 Steps in Standard Setting

A description of the steps typically followed in a standard-setting study is a useful introduction to not only the nature of the activities that comprise the process but to later descriptions of different types of methods. Hambleton (1998) presented a useful summary of the procedures generally conducted as part of a standard-setting study. These steps are summarized in Table 2.1, and descriptions of each step follow.

Step 1: Choose a panel. Since the establishment of passing standards may affect several groups of stakeholders, each of these groups should be represented on the panel. As an example, for educational tests these groups may include teachers, administrators, curriculum specialists, policy makers, and members of the public.

Step 2: Choose a method, prepare materials, and finalize agenda. The choice of standard-setting method is an extremely important step in the process. As will be outlined below, there are numerous methods from which to choose, and each has advantages and disadvantages which should be considered carefully before a selection is made. Once the method is chosen, training materials should be prepared that will facilitate the panelists' execution of required tasks. The agenda should allow ample

time for these tasks to be completed in as thorough a manner as possible (ideally, training materials will have been field-tested in order to obtain an estimate of the time needed to perform different steps within the procedure).

Table 2.1

Hambleton's (1998) Steps for Standard Setting

Step number	Step description
1	Choose a panel (large and representative of the stakeholders)
2	Choose a standard-setting method, prepare training materials, and finalize the meeting agenda
3	Prepare descriptions of the performance categories
4	Train the panelists to use the method (including practice in providing ratings)
5	Compile item ratings or other data from the panelists
6	Conduct a panel discussion, consider actual performance data, and provide feedback on inter-panelist and intra-panelist consistency
7	Compile item ratings a second time (may be followed by more discussion and feedback) [optional]
8	Compile panelist ratings and average to obtain the passing standard
9	Present consequences data to the panel [optional]
10	Revise, if necessary, and finalize the passing standard(s) [optional]; conduct a panelist evaluation of the process itself and their level of confidence in the resulting standard(s)
11	Compile technical documentation to support the validity of the passing standard(s)

Step 3: Prepare descriptions of the performance categories. Clear descriptions of the nature of candidate performance to be reflected in each category are an essential component of the standard-setting process. As noted earlier, the cut score that is the end-product of the standard-setting study is an operationalization of the performance standards; thus, the starting point (performance standards) must be clearly understood by panelists. The performance category descriptions may have been previously formulated at earlier meetings or by policy makers or a licensing agency, or may be drawn up as a preliminary part of the standard-setting study. In any case, it is important that panelists be encouraged to discuss the performance standards until they are clear in their own minds on what the differences in performance are.

Step 4: Train the panelists to use the method. A theoretically effective method is only as good as its practical implementation, and this implementation depends in large part on the quality of the training that panelists receive. It is important that panelists have a clear understanding of the steps involved in standard setting, gain a familiarity with the types of materials to be used in the process (i.e., text of items, scoring rubrics, rating forms), have a chance to practice making ratings, and understand the nature of any data they will be given during the process (i.e., examinee performance data or information on panelists' ratings). In addition, it is often helpful as a part of this step to have panelists take all or some of the items as part of a practice test. This often serves as a potent reminder of the true difficulty of the items, since viewing them in the absence of the scoring key and under timed conditions may give a much different impression than first perusing the items with the answer key available.

Step 5: Compile item ratings or other data from the panelists. In this step, panelists execute one of the main tasks by which a given standard-setting method is known—providing judgments. For example, in the Angoff method panelists may provide an estimate of the proportion of borderline candidates who would answer the item correctly. After panelists have completed their task, appropriate data is compiled. In the Angoff method, for example, a mean rating across panelists may be calculated for each item.

Step 6: Conduct a panel discussion; provide data and feedback. After data have been compiled from the tasks done by panelists in step 5, this information is often presented to group members for discussion. In addition to the item rating or other data noted above, information presented to panelists may include actual examinee performance data. Panelist-specific information may also be provided, such as indications of inconsistency within one panelist's ratings, and inconsistent panelists may be asked to explain their ratings (van der Linden, 1982). Overall, the group discussion that is conducted with the provided data as a focus is often beneficial in helping panelists' clarify their positions and, at times, to change them.

Step 7: Compile item ratings or other data a second time. After the initial panel discussion, Step 5 may be repeated, giving panelists a chance to revise their ratings. In addition, a second round of discussion may ensue. This step is optional, though such an iterative process is often recommended.

Step 8: Compile panelist ratings and average to obtain passing standard. The ratings compiled in step 7 are compiled in order to determine the group passing standard. In the Angoff method, for example, each panelist's item ratings are summed

to get a test cut score; then, these panelist cut scores are averaged to obtain a group passing standard.

Step 9: Present consequences data to panelists. In this optional step, data regarding the impact of these standards on the rate of examinee classifications may be provided to panelists. For example, panelists could be informed that the resulting passing standard results in only 20% of the candidates for certification being classified as “passing.”

Step 10: Revise and finalize standards; conduct evaluation. If consequences data has been presented to panelists in step 9, they may be allowed to revise their ratings given impact on examinee classifications. In all cases, regardless of whether step 9 has been executed, it is important to gather panelist feedback regarding their confidence in the process. A questionnaire is usually administered for this purpose.

Step 11: Compile technical documentation. It is essential for validation purposes to document the steps that were taken in the standard-setting process. Such documentation will serve as needed support for the validity of future interpretations made from test scores.

2.2 Classification Dimensions

A common dichotomy used to distinguish among standard-setting methods is that of test-centered methods versus examinee-centered methods (Cizek, 1996a; Jaeger, 1989; Kane, 1994). Test-centered standard-setting methods require panelists to make judgments about test content. During their review of test items, panelists provide

judgments regarding expected levels of performance on each item by examinees on the border between two levels of performance.

Examinee-centered standard-setting methods focus on examinees rather than on test items. Standard-setting panelists classify examinees into performance categories, such as “pass,” “fail,” or “borderline,” according to a process specified by the particular method. In the borderline group method, for example, standard-setting panelists identify borderline examinees on the basis of their knowledge of the examinees, after which the test scores for these borderline examinees are gathered and their median test score is typically used as the cut score. In the contrasting groups method, panelists identify two different groups of examinees, one whose members are clearly above a particular standard and another whose members are clearly below that standard. The cut score results from a contrasting of the test score distributions of these two groups.

In addition to the test centered/examinee centered distinction, however, there are other dimensions along which standard-setting methods may be classified. In fact, these dimensions are often necessary to fully understand the differences among the emerging methods of standard-setting that are described later in this review. Hambleton et al. (in press) outline the following six dimensions that may be used to differentiate standard-setting methods.

Dimension 1: Focus of Panelists’ Judgments. The panelists may be instructed to focus on one of four types of stimuli in order to make their judgments. The first type is tasks or item on the assessment, including scoring rubrics if applicable. The second is the examinees themselves. A third type is examinees’ responses to the tasks or items

on the assessment. The fourth type of stimulus is candidates' scores on those tasks or items.

Dimension 2: Panelists' Judgmental Task. The second dimension is linked to the first. Given the focus of the panelists' judgments, what is their task? First, if panelists are focused on items, they may be asked to estimate the performance of borderline examinees on those tasks. In the second case, where the focus is on examinees, panelists may be asked to sort those examinees into performance categories. Third, if examinee responses are the focus, panelists may be required to classify those responses into categories or determine which are characteristic of borderline examinees. And fourth, when panelists focus on scored performances, they may be asked to identify the performance categories into which those scored work samples should be sorted.

Dimension 3: Judgmental Process. The judgmental process may be characterized in several ways. Judgments may be made individually or in a group setting. And as discussed earlier, the types of feedback given may vary, and there may be a second round of ratings after the initial round.

Dimension 4: Composition and Size of Panel. The panel may be composed of different types of members, including experts or stakeholders. The panels may be homogenous or heterogeneous, and their size may vary as well.

Dimension 5: Validation of Resulting Passing Standards. The validity of the resulting passing standard must be supported by different types of evidence. Examples of evaluation criteria are discussed in section 2.5 of this review.

Dimension 6: Nature of the Assessment. An assessment may be characterized by several features. For example, the types of items comprising the assessment may

include multiple choice or constructed response. In addition, the assessment may be unidimensional or multidimensional. Scoring may be compensatory or conjunctive.

These six dimensions proposed by Hambleton et al. (in press) provide a flavor of the many ways in which standard-setting processes may vary. They serve as a useful introduction to the descriptions of more specific standard-setting methods that are presented in the following sections.

2.3 Traditional Standing-Setting Methods

There are several standard-setting methods that had been in primary use until research into new methods began in the last decade (and these methods still make up the bulk of those used operationally in licensure and certification settings, according to Meara, 2000). The traditional methods to be described in this chapter can be categorized most easily in terms of the traditional test- vs. examinee-centered dichotomy (Cizek, 1996a; Jaeger, 1989; Kane, 1994). Four of the methods are test-centered methods—the Angoff, Ebel, Nedelsky, and Jaeger methods. Two are examinee centered methods—the contrasting groups and borderline groups methods.

2.3.1 Angoff Method

What is now known as “the Angoff method” was first described by Angoff in his chapter “Scales, Norms, and Equivalent Scores,” in the second edition of Educational Measurement (Angoff, 1971), and was subsequently reprinted as Angoff (1984). This heavily cited introduction of the Angoff method is limited to two paragraphs, one of which is a footnote. The method described in the footnote requires

standard-setting panelists to review each multiple-choice test item and provide an estimate of the proportion of borderline examinees who would answer the item correctly. The method described in the main text itself is the simpler version (which Angoff attributed to Ledyard Tucker), in which the panelists merely decide whether the borderline examinee would answer the item correctly or not. In either case, the ratings for each panelist are summed across items, and these sums are averaged across panelists, to calculate the cut score. As such, the Angoff method is a test-centered approach to standard setting. Newer variations on this method use the term “modified Angoff” to reflect the addition of one or more features not present in the original formulation. These newer features include providing empirical item data to participants, encouraging discussions among panelists, and conducting several rounds of ratings to enable panelists to revise their estimates (Cizek & Fitzgerald, 1996; Mills, 1995).

Variations of the Angoff method are the most popular for setting standards on educational tests (Kane, 1994; Mehrens, 1995). In addition, three surveys have indicated that the modified Angoff method is the most commonly used method for licensure tests (Meara, 2000; Plake, 1998; Sireci & Biskin, 1992). Cizek (1996a) also observed that the Angoff method has been subjected to the most vigorous research and has been the most widely used. A review of research regarding different features of the Angoff method, such as the types of ratings made and what kind of information is provided to panelists, can be found in Pitoniak and Sireci (1999).

However, the Angoff method has been subjected to the criticism that the very task inherent in this method—evaluating the difficulty of test items—is too difficult for

panelists to accomplish in an accurate manner (Shepard, Glaser, Linn, & Bohrnstedt, 1993). Angoff (1988) also acknowledged that more attention should be paid to factors affecting the reliability of item judgments, and noted that lack of agreement in cut scores may stem from two factors. First, the panelists may not have a clear picture of the competency of the borderline examinee. Second, even if panelists did have a clear picture, they may not be able to accurately determine probabilities of correct responses to these items.

In one of the more widely known attacks on the Angoff method (in the context of the National Assessment of Educational Progress [NAEP]), panelists' systematic overestimation of performance on difficult items and underestimation on easy items was noted (Shepard, 1995). Researchers reviewing 1990 NAEP standard setting concluded that accurately estimating performance probabilities is an "unreasonable cognitive task" (Shepard et al., 1993, p. 72). However, several psychometricians and policy makers strongly defended the procedures that were followed. The National Assessment Governing Board (NAGB), which coordinated these standard-setting efforts, responded by stating the alternative methods suggested in the United States General Accounting Office (USGAO) report "appear naïve and unsupported by research evidence" (USGAO, 1993, p. 88). This position was articulated further by Kane (1995). In addition, Hambleton, Brennan, et al. (2000) recently presented a rebuttal to a critical summary of NAEP standard setting compiled by Pellegrino, Jones, & Mitchell (1999). Hambleton, Brennan, et al. concluded that the Pellegrino et al. report "presents a very one-sided and incomplete evaluation that is based largely on dated and second-hand evidence" (p. 6); they also presented a review of evidence that refutes the report and

supports the credibility of the Angoff standard-setting method as implemented for NAEP.

A fair amount of research has been conducted on the ability of panelists to provide accurate ratings. Several studies provided results that support the view that panelists are capable of providing accurate item ratings (Goodwin, 1999; Plake, Impara, & Irwin, 1999). In contrast, other studies have shown that panelists have difficulty with the task (Bejar, 1983; Impara & Plake, 1998). Thus, evidence for the accuracy of panelists' item ratings appears to be mixed. The degree to which panelists can accurately estimate the probability of an examinee getting an item correct depends on the training of the panelists, the type of empirical data they receive, and the difficulty levels of the items being rated. In particular, Kane et al. (1997) argued in their theoretical evaluation of the Angoff method that it does not contain adequate controls on the standard's being set too high. They suggest that items with extreme *p*-values be eliminated from the rating process so as not to bias the estimates, since as reported by Shepard (1995), panelists have a difficult time estimating borderline examinee performance on these items.

2.3.2 Ebel Method

In the Ebel (1972) method, standard-setting panelists make item-by-item judgments and classify items along two dimensions—difficulty and relevance. Then, for each combination of difficulty level and relevance level, panelists provide a judgment (via expected percent correct) as to how the borderline examinee will perform on the items contained within that combination. The cut score is obtained by

multiplying the number of test items in each cell by the percentage assigned by the panelist, summing those products, dividing by the total number of test items, and then averaging these scores across panelists. The Ebel method may be used for both dichotomous and polytomous items.

Berk (1986) questioned whether keeping the highly-correlated dimensions of difficulty and relevance distinct is too difficult a task for panelists. Cizek (1996a) pointed out that the Ebel method may prompt questions about the test construction process itself, since the method identifies items that are of questionable relevance. In addition, Cizek noted that requiring panelists to come up with item difficulty levels may not seem necessary since empirical item data are often available.

Perhaps for these reasons, the Ebel method is not one of the more frequently used standard-setting approaches (Meara, 2000; Sireci & Biskin, 1992). Similarly, in her review of standard-setting methods used for licensure and certification tests, Plake (1998) did not mention that any agencies used the Ebel method. A review of research did not reveal recent empirical investigations of this method.

2.3.3 Nedelsky Method

The Nedelsky (1954) method is a test-centered approach that involves raters making judgments about test items. In this method, panelists estimate for each item the number of distractors that they think the borderline examinee would be able to rule out as incorrect. The probability that the borderline examinee will answer the item correctly is the reciprocal of the number of distractors not ruled out. Then, as in the

other test-centered methods, the estimates are averaged across items for each panelist, and then averaged across panelists to yield a final cut score.

The Nedelsky method has been criticized for its tendency to produce inaccurate, usually low, cut scores (Shepard, 1980). Research has borne out these concerns (e.g., Chang, 1999; Melican, Mills, & Plake, 1989; Subkoviak, Kane, & Duncan, 1999). The Nedelsky method is not as widely used or researched as the Angoff method. Although it is the method used to set the passing scores on the National Optometry Licensing Examination, its use appears to have declined in recent years (Mills, 1995). Meara (2000) found only a handful of credentialing agencies that used the Nedelsky method.

2.3.4 Jaeger Method

Although the Jaeger method (Jaeger, 1982, 1989) is a test-centered method, it differs from those described above in that it deliberately takes into account the various constituents who may have a stake in the standard being set. This method focuses on whether panelists, via the use of a yes/no method, think the borderline examinee should be able to answer the item. In that sense, as Kane (1994) pointed out, the focus is shifted from estimating a probability for a hypothetical group of examinees to a more overtly value-laden judgment. The Jaeger method was originally formulated as an iterative process, which is now a common feature of other standard-setting methods (such as Angoff), as well.

Applied examples of the Jaeger method are hard to find. In their review of the professional licensure arena, neither Meara (2000) nor Sireci and Biskin (1992) listed the Jaeger method as one of the methods used by the participating organizations.

Similarly, Plake (1998) did not mention use of the Jaeger method in her review of standard-setting methods used by licensing and certification programs.

2.3.5 Borderline Group Method

In the borderline group method, standard-setting panelists identify borderline examinees on the basis of their knowledge of the examinees. The test scores for these borderline examinees are then gathered and their median test score is typically used as the cut score. Research related to this method will be described in the next section, since the studies that examined this method (borderline group) also looked at the next method to be described (contrasting groups).

2.3.6 Contrasting Groups Method

In the contrasting groups method, panelists are used to identify a group of examinees whose members are clearly above a particular standard and another group whose members are clearly below that standard. The test score distributions of these two groups are then contrasted to select the cut score. There are several variations of how to determine the cut score; Livingston and Zieky (1982) described smoothing the distribution and selecting the point at which 50% of the candidates were qualified. In a different approach, the test score that results in the fewest “false positive” errors (i.e., classifying a below-standard candidate as meeting the standard) and “false negative” errors (i.e., classifying an above-standard candidate as not meeting the standard) is selected as the cut score. Livingston and Zieky (1989) used logistic regression to find

the test score that minimized these two types of errors. This strategy was also used by Sireci, Rizavi, Dillingham, and Rodriguez (1999) and Sireci, Robin, and Patelis (1999).

Cizek and Husband (1997) used a Monte Carlo approach to analyze the effects of different population characteristics, sample strategies, sample size, and panelist error rates on the cut scores obtained using the contrasting-groups method. They noted four significant findings. Stable estimates of the cut score could be produced with a sample size as low as 100 candidates; negatively skewed and symmetric sampling strategies appear to work best; panelist error rates were not found to have a significant effect on the accuracy of the cut score estimation; and the accuracy of the cut score increased as the proportion of candidates classified as masters declined from 80% to 60%.

Livingston and Zieky (1989) compared the Angoff, Nedelsky, borderline group method, and contrasting group methods. They found that when the target populations included approximately equal numbers of students classified as masters and nonmasters, the borderline group and contrasting group methods produced similar results, but they differed when the proportions were not equal. In the latter case, the contrasting groups cut scores were biased in the direction of whichever group was smaller. This latter finding echoes the results found by Cizek and Husband (1997).

Giraud, Impara, and Buckendahl (2000) compared several different standard-setting approaches in a school-district setting. They looked at the borderline and contrasting groups methods, the yes/no version of the Angoff method, and two new methods—one based on course enrollment and one based on the expectations of experts. They found that the methods generally produced similar cut scores.

2.4 Newer Standing-Setting Methods

The standard-setting methods to be described next have been developed and researched within the past decade. Many, though not all, of these methods attempt to address the unique features of performance assessments and other polytomously-scored items. Others seek to avoid the necessity to provide item-level ratings such as those required by approaches like the Angoff method.

2.4.1 Cluster Analysis Method

The cluster analysis method (Sireci, 2001; Sireci, Robin, & Patelis, 1999) uses examinee response data (i.e., scored responses to test items) to form borderline or contrasting groups. In this method, which is appropriate for tests comprising dichotomous and/or polytomous items, examinees are compared with one another on the basis of their performance on individual items or groups of items. Test-takers who are most similar to one another with respect to test performance are grouped together into clusters. The standard-setting task is to arrange these clusters from lowest performing to highest performing (e.g., based on average test scores for examinees within each cluster) and then decide which clusters are to be used as borderline or contrasting groups. The advantage of this method over traditional borderline and contrasting groups methods is that expert panelists are not needed to identify students for the borderline and contrasting groups. These groups are “discovered” through cluster analysis of item score data.

Sireci, Robin, and Patelis (1999) applied this procedure to a statewide mathematics test that classified students into three achievement levels (intervention,

proficient, and excellence). The cut scores derived using cluster analysis were validated using students' final math course grades. They concluded the method was effective for facilitating the standard-setting process and suggested it could be used to provide supplementary information to panelists participating in a test-centered standard-setting study. In another study, Sireci (1995) explored the use of cluster analysis for setting standards on the writing skills component of the GED Tests, and found that the standards set using the cluster-analysis procedure were similar to those recommended by the GED Testing Service. These latter standards were recommended by setting the passing score for the adults who take this test at the 30th percentile of the high school senior norm group. In evaluating the cluster analysis procedure, Sireci (2001) concluded that it was useful for: (a) setting cut scores without employing panelists, (b) deriving profiles of test-takers' performance that could be used in judgmental policy capturing or dominant profile method studies (see sections 2.4.2 and 2.4.3, respectively, for descriptions of those methods), and (c) setting standards on multidimensional tests, such as those comprising various item types.

However, a study conducted by Meara (2000) raises questions about the cluster analysis method. He used the procedure with data from a test for which standards had been previously set with the Body of Work method (see section 2.4.7 for a description of that method). The standards yielded by cluster analysis were incongruent with the previous method's results, and with teacher ratings. The results of this study suggest that the cluster analysis method may not perform consistently with different types of score distributions.

2.4.2 Judgmental Policy Capturing Method

In the judgmental policy capturing (JPC) method, panelists review hypothetical score profiles across items composing a performance assessment, and assign each to a proficiency level. These data are then analyzed to determine each panelist's latent standard-setting policy. To obtain the group's latent standard-setting policy, a weighted average of the panelists' policies is calculated. The resulting policy may be one of three types: (a) compensatory, meaning that the total score is a weighted total of scores on individual exercises; (b) conjunctive, meaning that some of the exercises would have a minimum required level, or (c) a combination of compensatory and conjunctive. Although this method was designed for performance assessments, examinee profiles could be constructed by grouping multiple-choice items according to their content designations, and assigning content-specific sub-scores to examinees.

Jaeger (1995) described the use of the judgmental policy capturing method with a National Board for Professional Teaching Standards (NBPTS) performance assessment. He concluded the method is feasible, panelists are up to the task of providing ratings on numerous complex assessment components in a reasonable amount of time, and there is a high level of intrapanelist consistency in responses to the score profiles. However, he also noted the standards resulting from the judgmental policy capturing method were higher than those obtained using the extended Angoff method, and appeared to be too high. He suggested several modifications to the procedure that could ameliorate the setting of standards that are too high. These modifications included allowing panelists an opportunity to discuss initial judgments, giving them information regarding the impact of their recommendations, and instituting a second

round in which panelists could revise their judgments. He also hypothesized that these modifications would reduce the variability of panelists' ratings. However, Hambleton (1998) noted that finding statistical models that fit the panelists' ratings, and then explaining the overall process to panelists, are drawbacks to the JPC method.

2.4.3 Dominant Profile Method

In the dominant profile method (DPM), panelists review score profiles across different exercises in the assessment and attempt to come to a consensus on the policy to be used in setting a standard. As in the JPC method, the policy to be formulated may be compensatory, conjunctive, or a combination of both. Similar to JPC, DPM was designed for performance assessments, but could be used with multiple-choice exams, if sub-scores were derived across content areas.

Putnam, Pence, and Jaeger (1995) conducted an investigation of the dominant profile method (also using the NBPTS performance assessment). They had recognized the JPC method may be premature in its attempt to capture panelists' standard-setting policies, and would be better used as a tool for helping them to formulate these policies. Hambleton (1998) noted that this is indeed an advantage of the dominant profile method; it allows panelists to engage in extensive discussions in order to determine what they think is the best standard-setting policy. However, any remaining divergence of these policies makes it difficult to reconcile them into one group policy. This shortcoming was also noted by Plake, Hambleton, and Jaeger (1997).

2.4.4 Bookmark Method

In the bookmark method, also called the item mapping method, panelists review specially constructed booklets in which the test items are ordered according to their difficulty parameter as estimated with an IRT model (Lewis et al., 1996). Panelists also receive an item map, which lists items in the sequence of their location in the ordered booklet and indicates each item's position in the original test booklet. The map also contains the content area designations of the items.

Lewis et al. (1996) asked panelists to place a bookmark “between two items on the item map such that from [your] perspective, the items preceding the cut-line represent content that all proficient students should be likely to know and be able to do (with at least a 2/3 likelihood of knowing the correct response for multiple-choice items or of obtaining at least the given score point for constructed response items)” (p. 3). The cut score is set by looking at the point on the ability scale where the bookmark was placed. As a result, they noted, judgments are made at the level of the cut score, not the item, although all items are of course reviewed during the process. The cut score determined by bookmark placement is translated to a scale score for each panelist by taking the mean of the IRT item location values of the items immediately preceding and following the bookmark. The final cut score, in turn, is taken by calculating the mean or median of the panelists' scale score cut scores. The bookmark method also facilitates the creation of descriptions of what students know and can do in each performance category, since panelists are focusing on item content (rather than on item difficulty, which is the case in many standard-setting methods, including Angoff). Lewis et al.

noted that panelists are able to operationalize what they expect of students at each level in terms of content of the test, as opposed to in terms of an idealized curriculum.

Lewis et al. (1996) acknowledged several potential problems with the bookmark method. If the test does not contain items representing the full range of ability levels for which cut scores are being formulated, a floor or ceiling effect may occur. For example, if a test does not contain difficult items for the advanced student, the bookmark placement will not accurately reflect the content that this type of student knows and is able to do. This occurred in their study for several committees who set the advanced cut score within the last 10 items in the test. Also, in terms of the creation of item descriptions, the authors found that panelists sometimes became confused about which items truly represented those that a student at a given level should be likely to know and be able to do. In addition, Mitzel, Lewis, Patz and Green (2001) acknowledged that research needs to be done into issues such as the impact of the ordering of items due to different measurement models and the density of items at certain points on the difficulty scale.

In reviewing the bookmark method, Hambleton (1998) and Hambleton, Jaeger, et al. (2000) noted that research needs to be done on the effect that the ability level chosen has on the resulting passing standard. For example, how would using a 67% cut-off instead of a 50% cut-off affect the standard? Other research suggested by Hambleton, Jaeger, et al. includes investigation of whether rating both multiple-choice and open-response items is problematic; i.e., whether the open-ended items have a greater impact on the passing score than they do on the overall test score.

Reckase and Bay (1999) also noted a potential problem with the bookmark method. They noted that although the bookmark method should result in a cut score similar to that yielded by the Angoff approach, it uses much less information in that theta estimates for only two items are, in theory, used to determine the cut score. This, they pointed out, could produce estimates of standards with larger standard errors than the Angoff approach, in which panelists estimate cut scores for all items in the test. Reckase and Bay suggested one way of overcoming this problem would be to set cut-points on multiple subsets of the items in the test and average the results from the subsets, which would allow for a better estimation of the standard error of the estimates produced by the bookmark method.

On the positive side, the bookmark approach appears to be viewed favorably by panelists, who feel confident about the standards that were set using the method. Panelists in the Lewis et al. (1996) study reported they experienced the technique as being “rational, interesting, and professionally enriching” (p. 8). Panelists who participated in a bookmark approach to standard-setting for statewide student assessment in Wisconsin (State of Wisconsin, 1997) similarly felt positive about their experience.

2.4.5 Extended Angoff Method

The extended Angoff method is a generalization of the Angoff procedure described earlier to tests that include polytomously-scored items. This method requires panelists to provide an estimate of the expected score a borderline test-taker would obtain on a polytomous item. For example, if a calculus problem were scored on a ten-

point scale, panelists would review the item and the scoring rubric and then provide their best estimate of the score a borderline student would receive on the item.

Hambleton and Plake (1995) applied several standard-setting methods to the certification exams of the National Board of Professional Teaching Standards (NBPTS), including the extended Angoff method. When using the extended Angoff method, panelists estimated the scores borderline candidates would get on each of the three dimensions used to score each performance task. These estimates were summed to derive the expected score for the borderline candidate on each exercise. Panelists were also given an opportunity to suggest weights to use in combining scores across items.

In her critique of the extended Angoff method, Plake (1995) noted that the extended Angoff method appears to be the easiest to administer, and speculates that it would yield more replicable results. However, as Hambleton and Plake (1995) acknowledged, the extended Angoff method fails to take into account the underlying decision rule of the panelists. The Angoff method is a fully compensatory model, whereas panelists appeared to want a conjunctive model, in which candidates must pass certain exercises in order to be certified. The analysis of questionnaire data from the panelists revealed a discrepancy between (a) the high degree of confidence the panelists felt in the standard (which was set using a compensatory model), and (b) the fact that the panelists theoretically viewed the conjunctive model as most appropriate. This disparity troubled the authors, and they concluded the standard that was ultimately set was not solidly in line with the panelists' preferences.

2.4.6 Analytic Judgment Method

The analytic judgment method is specifically designed for tests that include polytomously-scored performance tasks (Plake & Hambleton, 2001). In this method, a carefully chosen subset of test booklets from real test-takers is used for analysis. All booklets must be previously scored, but these scores are not revealed to the panelists. The booklets are selected to represent specific points along the composite test score scale and along individual item score distributions. The “analytic” feature of this method is that panelists’ ratings are based on components of the test, rather than on the entire test, as in holistic methods.

Although there are several variations of the analytic judgment method, Plake and Hambleton (2000, 2001) found that a sorting procedure works well, and is relatively simpler than other methods. In this variation, for each section of the test, panelists are asked to review a subset of student papers and sort them into a number of pre-specified achievement categories. Panelists who are teachers like the method because they are more comfortable sorting student papers into ordered performance categories than they are providing Angoff-type estimates, since the former task is a common one for teachers. Plake and Hambleton applied the procedure to a NAEP science test, which comprised four achievement categories. Once the sorting task was completed, the panelists were asked to sort papers within each category into two or three more sub-groups (e.g., low, medium, high). Panelists then discussed their individual assortments and made changes, if they desired.

The end result of this sorting procedure was an ordinal grouping of student papers for each panelist. To derive cut scores from the panelists’ ratings, the “piles”

relevant to each standard were identified, and the average test scores in those piles were calculated. Cut scores are derived by summing the section scores for borderline test-takers across all sections of the test. (Two different data analysis strategies were used—the boundary paper method and cubic regression models. The models were judged to provide similar cut scores.)

Recent applications of the analytic judgment method, including Plake and Hambleton (2000, 2001), suggest the procedure works well with tests comprising both multiple-choice and free-response items. For example, Buckendahl, Plake, and Impara (1999) conducted a study in which they used both a modified Angoff method (for multiple-choice items) and an analytic judgment method (for free-response items) in a school-district setting. To make the procedure more practical, both the panel and the assessment were subdivided. That is, no member of the panel reviewed all parts of the test, but there was overlap among the test parts to evaluate consistency among the panelists and parts. The authors concluded that this strategy appeared particularly useful for tests comprising both multiple-choice and constructed-response items on which standards need to be set quickly.

2.4.7 Body of Work Method

In the body of work method, also termed the holistic or booklet method, panelists review the complete work of a student, over all of the tasks in the assessment, and decide which booklets are most likely to represent the work of borderline test-takers. Thus, this method is more holistic in scope than the analytic judgment method. Kingston, Kahl, Sweeney, and Bay (2001) reviewed results from implementation of this

method in several statewide student assessment programs. They concluded that although this method, which utilizes a task similar to that which educators are accustomed to doing—reviewing a rich body of student work—is promising, more work is needed to explore why it often produces higher cut scores than other method.

Hambleton (1998) noted that one advantage of this method is that it allows panelists to provide judgments about the overall performance of a test-taker rather than focusing on the performance of individual items. More research is needed to study the strengths and limitations of the method, as well as its utility in comparison to the analytic judgment and Item Cluster methods.

2.4.8 Direct Consensus Method

The Direct Consensus method, one of the two approaches that were implemented in the current study, is based on a desire to streamline the standard-setting process. Many of the standard-setting methods, both new and old, arrive at passing standards via what must seem to panelists as rather convoluted procedures. For example, with item-level methods such as the Angoff method, panelists' item-level ratings are averaged, and then the item-level averages are summed to arrive at a passing standard. Even when panelists understand the calculations, they often fail to see how the procedures carried out can lead to a defensible or sensible passing standard. Most of the other methods suffer from the same flaw. The procedures for arriving at the final passing standards seem mysterious.

In the Direct Consensus method, panelists set passing standards directly based on a consideration of the descriptions of the standards of performance associated with

each cut score, a consideration of previous exams and the corresponding standards, the content of the current exam and its scoring rubrics, any statistical data that may be available, sample examinee constructed responses (as applicable), and more. The process involves using a facilitator to engage panelists in a discussion of all of the available information and attempt to reach consensus on the resulting passing standards.

The direct part of the method is that panelists work with the actual exam scale; the consensus part is that the goal is to have the panel arrive at passing standards that they can agree upon (as a last resort, the mean of their recommended passing standards can be used). An advantage of this method is that it is faster than other methods, because item level ratings are not provided, and panelists do not need to sort through rather large sets of student papers. However, panelists still must be familiarized with the test and scoring rubrics.

The Direct Consensus method is a new approach that was recently implemented by Sireci, Hambleton, et al. (2000) in a standard-setting study within a certification context. In this study, two different panels used both the Direct Consensus method and the Angoff method. One of the panels reached consensus on the standard, while in the other panel the score-averaging technique needed to be used. Most importantly, the two panels' passing scores were within one point of each other. This is in contrast to the Angoff method, where there was a three-point difference between panels. Also of note is the fact that the Direct Consensus method took slightly less time to implement than the Angoff method.

In terms of participant feedback, the Direct Consensus method was seen by panelists as an appropriate standard-setting method. All panelists liked one significant

feature of the Direct Consensus method—that panelists have direct control over the final standard (in contrast to other methods, where panelists may not be aware of the final standard and/or be able to adjust their ratings once the group standard is known).

2.4.9 Item Cluster Method

The Item Cluster method, the second of the two methods implemented in the current study, involves dividing an exam into homogeneous clusters of items according to their content areas and then presenting panelists with approximately 20 real or hypothetical patterns of student responses to the questions in each cluster. When examinees get a multiple-choice question incorrect, the panelists are informed of the distractors chosen by the examinees. This information along with information about the questions answered correctly can be used in judging the quality of an examinee’s work. For example, with a seven-Item Cluster of multiple-choice questions, a response pattern might look like the following: b 1 1 a c 1 1. Items 2, 3, 6, and 7, which have values of “1” listed, were answered correctly. Answers items 1, 4, and 5 were incorrect, and the distractors chosen (b, a, and c, respectively) are listed. For any constructed-response questions, panelists would see actual student work.

Panelists assign these student response patterns into one of six categories, ranging from 1 (hopeless) to 6 (exceptional). After completing their initial ratings, panelists meet to discuss their ratings. Then, following a discussion, panelists have a final opportunity to reclassify the student response patterns. This process is repeated for each cluster of items. In the Item Cluster method, arriving at a final set of standards can be handled in one of several ways: looking at the mean scores of student response

patterns assigned by panelists to the borderline categories (boundary method), fitting a non-linear regression line to the mean scores of student response patterns assigned to each of the six performance categories, or by determining an equating relationship between response pattern scores and assigned performance categories.

The advantages of the Item Cluster method are that it (a) can handle both multiple-choice and performance tasks in the same test, (b) allows panelists to consider the actual performance of students, but in small enough chunks that the examinee patterns of rights and wrongs, and actual work on the constructed response questions, can be meaningfully judged holistically, and (c) the method is focused on actual student work—something that panelists often say they want to consider in setting standards.

The Item Cluster method was one of the two methods used in the Mills et al. (2000) CPA Exam study (the other approach used was the Angoff method). As noted by Mills et al., the Item Cluster method can be seen as a hybrid approach incorporating aspects of both examinee-centered and test-centered methods. In that study, the cut scores yielded by the Item Cluster method were more consistent across panels than those obtained with the Angoff method (though this may have been confounded with a facilitator effect). The cut scores, while lower than those set using the Angoff method, were more consistent with the cut scores that resulted when the Beuk (1984) compromise method was utilized. Panelists also felt more positively about the Item Cluster method than about the Angoff method.

The current study allowed for further investigation of this method. One focus of the study was that greater attention was paid to the training given to panelists, since many panelists in the Mills et al. (200) study indicated that they would have liked more

extensive training in the method. In addition, both the training and evaluation components addressed the extent to which panelists consider the full pattern of examinee responses when assigning a rating to them. It appeared conceivable in the previous study that panelists tended to focus more on simply whether the examinee selected the correct or incorrect response option than on which incorrect response was selected when the examinee got the item wrong. One of the intended benefits of the Item Cluster method is that attention is given to more aspects of an examinee's performance, and it was important to monitor the extent to which this goal is being achieved.

Table 2.2 below contrasts the two methods used in the current study. Characteristics of the Direct Consensus method and the Item Cluster method are summarized along several dimensions in order to highlight the similarities and dissimilarities between the approaches.

Table 2.2

Direct Consensus and Item Cluster Methods: Similarities And Differences^a

Feature of method	Angoff method	Direct Consensus method	Item Cluster method
Panelists are familiarized with exam purpose	Yes	Yes	Yes
Panelists are familiarized with exam content	Yes	Yes	Yes
Panelists take exam (or set of exam items)	Yes	Yes	Yes
Panelists discuss just qualified candidate	Yes	Yes	Yes
Items are grouped into clusters	No	Yes	Yes
Panelists predict whether just qualified candidate will answer item correctly	Yes	No	No
Panelists review samples of item responses (patterns of responses to objective items, and samples of essays) and provide ratings of them	No	No for objective items, Yes for essays	Yes
Panelists discuss item ratings	Yes	No	Yes
Item statistical information can be introduced to inform panelists' discussions	Yes	Yes	Yes
Panelists discuss average passing score	Sometimes	Yes	No ^b
Panelists encouraged to reach consensus regarding passing score	No	Yes	No ^b
Panelists can change their individual passing score	Sometimes	Yes	No
Panelists leave the meeting knowing the recommended passing score	Usually	Yes	No ^b

Note. From Sireci, Hambleton, et al. (2000). Adapted with permission.

^aThe Angoff method, one of the most widely-used standard-setting methods (Kane, 1994; Mehrens, 1995) is also given for comparison purposes.

^aAlthough these features were not implemented in this study, they could be incorporated into the Item Cluster method if time and resources were allocated for them.

2.5 Evaluation Criteria

Once a standard-setting study has been conducted using a method such as the ones just described, the important task of evaluating the process must begin. As noted earlier, there are no absolute criteria against which standards can be validated and there are no perfect criteria for evaluating different standard-setting studies (Kane, 1994, 2001). However, the absence of absolute criteria does not excuse a testing agency from providing evidence that the standards are reasonable and appropriate, nor does it mean that one standard-setting method is as good as another (Hambleton, 1998; Jaeger, 1991; Linn, 1998).

Several sets of guidelines and recommendations for carrying out a standard-setting study appear in the literature (Cizek, 1993, 1996a, 1996b; Hambleton, 1998; Hambleton & Powell, 1983; Jaeger, 1991; Livingston & Zieky, 1982; Norcini & Shea, 1997; Plake, 1997). In addition, the Standards for Educational and Psychological Testing (AERA et al., 1999) stipulate several recommendations for conducting and evaluating standard-setting studies. In general, these guidelines discuss the need for carefully designing, conducting, evaluating, and documenting standard-setting studies. The degree to which such guidelines have been followed are often used (by both the courts and psychometricians) as criteria for evaluating the validity of examinee classifications based on standards (Sireci & Green, 2000).

Kane (1994, 2001) discussed three categories of evidence that can be used to support the validity of standards: (1) procedural, (2) internal, and (3) external. Table 2.3 summarizes the different sources of standard-setting validity evidence using these three broad categories. Although other authors have grouped their evaluation criteria in

different ways, in general their concepts can be contained within this three-pronged structure. For a review of evaluation guidelines grouped by author, the reader is referred to Plake (1997).

2.6 Summary

It is clear from the literature review presented above that standard-setting research has proliferated in the past few decades. Such research is valuable. As Norcini and Shea (1997) noted, it is crucial that any standard-setting method “be supported by a body of research, preferably published, that rules out threats to credibility and establishes that the standard has reasonable properties” (p. 45).

Among the most important data that Norcini and Shea (1997) suggest be gathered are two types relevant to the current study. The first is the comparison of a given method with competing methods. In the current study, this was accomplished by using both the Direct Consensus and the Item Cluster methods on the same set of test items. The second is establishing that a given method yields a reproducible standard. This criteria was investigated in the current study by using two panels for each method.

The empirical evidence to be provided by the current study thus allowed for numerous informative analyses. The evaluation of the resulting cut scores in the context of the validity criteria outlined above is an important step in the establishment of these two new methods as credible options for standard setting.

Table 2.3

Summary of Criteria for Evaluating Standard-Setting Procedures

Evaluation criterion	Description	Sources
<u>Procedural</u>		
Explicitness	The degree to which the standard-setting process was clearly and explicitly defined before implementation	van der Linden (1995)
Practicability	The ease of implementation of the procedures and data analysis, and the degree to which procedures are credible and interpretable to laypeople	Berk (1986)
Implementation of procedures	The degree to which the following procedures were systematic and thorough: selection and training of panelists, definition of the performance standard, and data collection	Kane (1994, 2001)
Panelist feedback	The extent to which panelists feel comfortable with the process and with the cut score	Kane (1994, 2001)
Documentation	The extent to which features of the study are reviewed and documented for evaluation purposes	Cizek (1996b); Hambleton (1998); Mehrens (1995)
<u>Internal</u>		
Consistency within method	The precision of the estimate of the cut score, or the extent to which same cut score would be obtained if method were replicated	Kane (1994, 2001); Cizek (1996b); van der Linden (1995)
Intrapanelist consistency	The degree to which a panelist is able to provide ratings that are consistent with the empirical item difficulties, and the degree to which ratings change across rounds	van der Linden (1982); Cizek (1996b); Berk (1996)
Interpanelist consistency	The consistency of item ratings and cut scores across panelists; includes "caution indices," whereby panelists are flagged whose ratings are inconsistent with the majority	Jaeger (1988, 1991); Cizek (1996b); Berk (1996)
Other measures	The consistency of cut scores across item types, content areas, and cognitive processes	Kane (1995)

(continued on next page)

Table 2.3 (continued)

Evaluation criterion	Description	Sources
<u>External</u>		
Comparisons to other standard-setting methods	The consistency of cut scores across replications with other standard-setting methods	Kane (1994, 2001)
Comparisons to other sources of information	The relationship between decisions made using the test to other criteria (e.g., grades, performance on a similar test, etc.)	Berk (1996); Giraud et al. (2000); Kane (1994, 2001); Shepard et al. (1993)
Reasonableness of cut scores	The extent to which the resulting cut scores are feasible or realistic, including impact on pass rates	van der Linden (1995); Kane (1998)

Note. From Sireci, Pitoniak, Meara, & Hambleton, 2000. Adapted with permission.

CHAPTER 3
METHODOLOGY

In this chapter, the methodology for the study is presented. Four major sections are included: study design, panelists, test items, and meeting procedures.

3.1 Study Design

There were two parallel panels in this study. Each panel used two different methods (one in the first session and one in the second session) and rated two different test forms. The order of the methods was counterbalanced across panels. The experimental design is shown in Table 3.1.

Table 3.1

Experimental Design

Panel	Session	
	Morning	Afternoon
A	Item Cluster (test form 1)	Direct Consensus (test form 2)
B	Direct Consensus (test form 1)	Item Cluster (test form 2)

3.2 Panelists

Panelists were recruited by the American Institute of Certified Public Accountants (AICPA). Notices that describe the study and solicit volunteers were sent to CPA firms, and CPAs who participated in a recent practice analysis survey were also

contacted to solicit their participation in the study. To be eligible for participation, CPAs must have practiced in a public accounting firm for a minimum of three and a maximum of seven years and must also be actively supervising new CPAs. Panelists received continuing professional education credits for their participation.

Upon arrival at the session, demographic information was collected for each panelist. This information included gender, years of experience, size of firm, number of new CPAs supervised, and number of years spent supervising new CPAs. During the orientation session, panelists were divided into two groups. Group assignments were balanced as much as possible in terms of the demographic characteristics noted above, with the following results. Gender was balanced perfectly across panels, with each panel having four men and four women. The panels were very closely matched in terms of number of years of experience. For Panel A experience ranged from 3-1/2 to 21 years, with a mean of 9.3; for Panel B experience ranged from 3 to 22 years, with a mean of 9.3. The level of experience possessed by the panelists is thus higher than the range originally desired (from 3 to 7 years).

Presented in Table 3.2 is information related to the number of new CPAs (with two years' experience or less) that each panelist supervised within the past two years, and the types of firms in which the panelists work. Of note is the fact that one panelist on each panel had not supervised new CPAs in the past two years, which means that these panelists did not meet one criterion for being a qualified participant. However, one of these panelists noted that she did supervise entry-level accountants, though they weren't yet CPAs.

Table 3.2

Characteristics of Panel Members

Characteristic	Panel	
	A	B
Number of new CPAs supervised in past two years		
None	1	1
1–3	2	3
4–10	1	3
11 or more	3	1
No response	1	0
Firm type		
Local	2	0
Regional	3	3
National	1	0
Big Five/International	1	2
AICPA	1	2
No response	0	1

3.3 Test Items

The Uniform CPA Examination is administered twice a year by the AICPA. The purpose of the exam is to “provide reasonable assurance to boards of accountancy that candidates passing the Uniform CPA Examination possess the level of technical knowledge and skills necessary for initial licensure to protect the public interest” (Board

of Examiners, 1996, p. 13). Four areas are covered by the CPA Exam: (1) Auditing (Audit); (2) Financial Accounting & Reporting (FARE); (3) Business Law and Professional Responsibilities (LPR); and (4) Accounting & Reporting—Taxation, Managerial, and Governmental and Not-for-Profit Organizations (ARE). Separate sub-scores are reported for each section; as a result, separate cut scores are also set for each section. For this study, standard setting was done for only one of the areas—the FARE section of the exam, administered in May 1998. This exam was also one of the two used in the Mills et al. (2000) study.

Three different assessment formats are used within the FARE section of the exam: (1) four-option multiple-choice questions, (2) other objective answer format (OOAF), and (3) essay question or problem format. The full-length version of the FARE section administered in May 1998 contained 60 multiple-choice questions, two OOAFs, and two essays. For the current study, a subset of these items was used; each test form contained 35 multiple-choice questions and one OOAF. Ten of the multiple-choice items overlapped across the test forms. An essay was not included due to time constraints.

From the overall group of items selected for the study, small groups or clusters of items were formed. For multiple-choice item clusters, 7 to 10 items were presented; the same clustering of items by content area were used as in the Mills et al. (2000) study. The one OOAF for each test form was presented as a separate cluster; the OOAF for test form 1 contained 11 items, and the OOAF for test form 2 contained 10 items. OOAFs are sets of objectively-scored questions based on a common stimulus or problem. Response formats include matching, classification, multiple yes/no, and

numeric constructed responses; thus the number of response choices can vary widely for items within OOAFs. The same clusters of items were used for both standard-setting methods (Direct Consensus and Item Cluster) for each of the two test forms.

3.4 Meeting Procedures

The meeting began with all of the panelists together for an orientation and training session. Thereafter, the panelists were split into two groups for the remainder of the day. The meeting agenda is presented in Table 3.3. The methodology of each component of the meeting follows.

3.4.1 Orientation and Training

Following a welcome and introductions, a general orientation was conducted regarding the setting of passing standards on exams. As noted in the standard-setting manual for the previous study (AICPA, 1999), “the main points to be made are that (1) a performance standard will be set on one section of CPA exam using the professional judgments of practicing CPAs, and (2) the performance standard will not be set to establish a particular passing (or failing) rate, but to ensure that the public is protected from substandard CPA work. The ultimate goal is to establish performance standards on the . . . CPA exam sections that are high enough to ensure that only competent CPA candidates are licensed, and not so high that many competent practitioners are barred from becoming CPAs” (pp. 2-3).

Table 3.3

Meeting Agenda

<u>Time Period</u>	<u>Activity</u>
8:30–9:00 A.M.	Panelists arrive; have coffee; fill out biographical form
9:00–9:10 A.M.	Introduction and orientation
9:10–9:30 A.M.	Panelists answer subset of items, followed by self-scoring of items
9:30–9:45 A.M.	Defining the minimally competent CPA
9:45 A.M.	Group splits into two panels
<u>Panel A</u>	
9:45–10:15 A.M.	Training for Item Cluster Method
10:15–10:30 A.M.	Coffee break
10:30 A.M.–1:00 P.M.	Conduct Item Cluster Method
1:00–1:45 P.M.	Lunch
1:45–4:30 P.M.	Training and Conduct of Direct Consensus Method
<u>Panel B</u>	
9:45–10:15 A.M.	Training for Direct Consensus Method
10:15–10:30 A.M.	Coffee break
10:30 A.M.–12:30 P.M.	Conduct Direct Consensus Method
12:30–1:15 P.M.	Lunch
1:15–4:30 P.M.	Training and Conduct of Item Cluster Method
<u>Panels Reconvene</u>	
	(if time/pacing allows; otherwise these activities will be conducted separately)
4:30–5:00 P.M.	Collect evaluation information; conduct discussion to gather further feedback on methods

Following the general orientation, panelists were given a short (ten-item) practice test under exam-like conditions. The purpose of the practice exam was to remind panelists of the experience of taking an examination under timed conditions. Next, a discussion was led regarding the nature of the minimally competent CPA. A description of the minimally competent CPA that had been previously adopted by the AICPA Board of Examiners was distributed for review and discussion. The goal of this part of the meeting was to ensure that panelists clearly understand the level of knowledge and skills that the minimally competent CPA has in preparation for the review of actual examination material and setting of standards.

After the general orientation and training sessions was completed, the panelists were split into two groups. Method-specific training and execution of the first standard-setting method then began. The methodology for each of the two methods will be described next.

3.4.2 Direct Consensus Method

In this method, test items are grouped into clusters of approximately seven to ten items each according to their content specifications (the same item clusters were used as were formed for the Mills et al., 2000, study for the Item Cluster method). For each cluster, panelists were asked to individually indicate on a rating form the number of items that they think the borderline candidate would get correct. A sample rating form is presented in Appendix A.

After rating of all clusters was completed, the ratings for each panelist for each cluster were placed into spreadsheet form and projected onto screen visible to all

panelists. The preliminary cut score (i.e., passing score, or the sum of the mean group scores for each cluster) was also shown on this spreadsheet. Table 3.4 shows a sample spreadsheet displaying panelist data. The “% correct” values represent the panelist mean divided by the number of items for the cluster or test form as a whole, and thus indicate what percentage of items the panelists think the borderline examinee should be able to answer correctly. One additional feature—the provision of actual examinee performance data to panelists—is optional in this method. These data were presented to panelists in this study after round one ratings to ensure consistency with the Item Cluster method, in which performance data are provided.

Table 3.4

Sample Panelist Data Display For the Direct Consensus Method

Panelist	Summary of Panelists' Passing Scores					
	Cluster					Passing Score
	1	2	3	4	5	
1	7	6	5	7	6	31
2	7	6	6	8	5	32
3	6	7	7	8	7	35
4	7	6	6	7	7	33
5	7	5	6	7	7	32
6	7	7	6	8	7	35
7	7	7	5	7	8	34
8	6	7	6	7	6	32
Panelist Mean	6.8	6.4	5.9	7.4	6.6	33.0
Number of Items	9	8	8	10	11	46
% Correct	0.75	0.80	0.73	0.74	0.60	0.72

A group discussion then ensued in which panelists explained the rationale for their ratings for each cluster. After the discussion, panelists provided a second round of cluster ratings, which in turn affected the projected cut score. Both panelist ratings and the projected cut score were displayed to panelists after the ratings were revised. Upon viewing these ratings, panelists were given an additional chance to change their total cut score if they felt that the sum of their cluster scores did not reflect their overall sense of how well examinees should be expected to perform; this was termed the global modification step. Next, the group discussed how viable they saw the group mean of the total cut score to be, and adjust it as they saw fit. If the group arrived at a consensus regarding a score, it was deemed the final group standard. If consensus was not achieved, the mean score across panelists was used as the final group standard.

3.4.3 Item Cluster Method

Responses from 17 to 20 examinees to each item cluster were presented to panelists. The same responses as those used in the Mills et al. (2000) study were used. Those responses were selected from a performance sample of 1,000 examinees to represent a distribution of scores on each cluster. The distribution of candidate response strings for the multiple-choice items and OOAFs is shown in Table 3.5. When there were no candidates with cluster scores in a particular score category, candidate response strings in other categories were over-sampled. Candidate response strings were selected in order to reflect a variety of response patterns. That is, responses from the same score category reflected different items as correct and incorrect.

The first-round data presented to panelists included not only an indication of whether the examinee got the item correct, but if he or she got it incorrect, the identity of the distractor that was chosen. Examinee responses were presented in order from lowest to highest score on that cluster.

Table 3.5

Distribution of Candidate Response Strings for Item Cluster Method

Cluster score (percent correct)	Number of candidate response strings
0 – 24.99	1
25.00 – 39.99	2
40.00 – 48.99	3
49.00 – 58.99	4
59.00 – 68.99	4
69.00 – 78.99	3
79.00 – 88.99	2
89.00 – 100	1

The task of the panelist was to review the examinee profiles on the cluster of items and rate each examinee’s performance on a six-point scale ranging from “hopeless” to “exceptional.” The rating scale is presented in Table 3.6, and a sample rating form is presented in Appendix A. Panelists were encouraged to look at the entire pattern of responses, including incorrect answers, before assigning the examinee to a

category. In addition, panelists were reminded to look at the number of response options for each item; while multiple-choice items have four response options, OoAFs may have, for example, 20 response options.

Table 3.6

Rating Scale for Item Cluster Method

Rating	Performance category
1	Hopeless
2	Failing
3	Just Below Borderline
4	Just Above Borderline
5	Solid/Strong
6	Exceptional

After the panelists completed their individual ratings for each cluster, a summary of the panelists' ratings (the number of panelists who placed the examinee into each of the six categories) was provided, along with performance data. A group discussion was conducted to review examinee ratings that were moderately to widely discrepant. Following the discussion, the panelists provided ratings for each cluster again, which gave them an opportunity to change their rating if the discussion led to a change in their evaluation of any of the candidates.

3.4.4 Collect Evaluation Information

Evaluation surveys were distributed to panelists. As with the Mills et al. (2000) study, there were four sections to the survey. The first contained general questions about the nature of the discussion of the minimally competent CPA and other orientation topics. The second section contained questions about the Direct Consensus method, and the third contained questions about the Item Cluster method. Within each of the method-specific sections, panelists were asked about the training for that procedure, factors that influenced their selection of a passing standard, and other questions designed to elicit their views about that method. The fourth section contained general questions and asked for any additional comments panelists wished to provide.

The format and content of the evaluation survey mirrored that from the Mills et al. (2000) study, though minor modifications were made (for example, to address in greater depth the degree to which panelists utilized all of the response information given to them in the Item Cluster method). The evaluation survey is presented in Appendix B.

CHAPTER 4

RESULTS

In this chapter, the results of the study are presented. Method-specific results are presented first, followed by additional summary information.

4.1 Direct Consensus Method

The aim of the Direct Consensus method is, as its name suggests, for panelists to come to a consensus on the final cut score. Both panels in this study were able to come to consensus. Panel B, which used the method in the morning, set a cut score of 34 on Test Form 1, which contained 46 items. In the afternoon session, Panel A set a cut score of 30 on Test Form 2, which contained 45 items.

4.1.1 Detailed Panelist Rating Information

While individual panelist cut scores were not used to calculate the final cut scores, information about these ratings is presented in Tables 4.1 and 4.2 in order to provide a complete picture of the nature of the ratings provided (for all analyses of the Direct Consensus method, results are presented for Panel B first, since they were the first panel to use this method on the day of the study). In addition to the individual panelist cut scores, the nature of the changes made by each panelist are made explicit in these tables. For Panel B (morning) only two panelists made changes to the cluster ratings that affected their cut score from round one to round two. However, six panelists made a change to their round two test form cut score to adjust it in the global modification step. For Panel A (afternoon), only three panelists made changes to the

cluster ratings that affected their cut score from round one to round two. Five panelists made a change to their round two test form cut score to adjust it in the global modification step. Thus the consensus nature of the Direct Consensus method appeared to have an effect even before the discussion took place during which the panel came to agreement on the final cut score.

Also of interest is information related to the spread of ratings for each of the five clusters. As shown in Table 4.3, for Panel B (morning), there were three instances of a one-point spread between the minimum and maximum, five instances of a two-point spread, and two instances of a three-point spread. For Panel A (afternoon, see Table 4.4), there was less variability in the ratings. Only once were the minimum and maximum rating more than one point apart—round one, cluster 5, where there was a two-point spread.

Analyses were also done to examine the degree of relationship between each individual panelist's ratings and empirical data provided to them. Panelists were provided with *p*-values, or the percentage of examinees at the operational administration that got the item correct, after the first round of ratings. Tables 4.5 and 4.6 present the correlations between the mean *p*-value for the item cluster and panelist ratings, the latter represented by the percentage of items that panelists judged a borderline candidate as needing to answer correctly. The four correlations are all moderate. However, the meaningfulness of these correlations is limited, since each is based on only five sets of data points, all of which are restricted in range. Scatterplots illustrating the location of the data points used in these analyses are presented in Figures 4.1 and 4.2.

Table 4.1

Panelist Cut Scores Across Rounds for Direct Consensus Method: Panel B (Morning)

Panelist	Round		Modified total score	Change: round		Change: round two to modified
	one	two		round one to two	two to modified	
B1	32	32	33	0	0	1
B2	35	36	34	1	1	-2
B3	33	33	34	0	0	1
B4	32	33	34	1	1	1
B5	32	32	33	0	0	1
B6	35	35	35	0	0	0
B7	31	31	33	0	0	2
B8	34	34	34	0	0	0
Panel mean	33.0	33.3	33.8	0.25	0.25	0.50
Consensus-based cut score	--	--	34	--	--	--

Table 4.2

Panelist Cut Scores Across Rounds for Direct Consensus Method: Panel A (Afternoon)

Panelist	Round		Modified total score	Change: round	
	one	two		round one to two	round two to modified
A1	30	30	30	0	0
A2	27	28	29	1	1
A3	28	28	30	0	2
A4	27	28	29	1	1
A5	31	31	30	0	-1
A7	29	29	30	0	1
A8	29	30	30	1	0
Panel mean	28.7	29.1	29.7	0.43	0.57
Consensus-based cut score	--	--	30	--	--

Table 4.3

Descriptive Statistics for Direct Consensus Method Ratings: Panel B (Morning)

Cluster	Number of items	Round one				Round two			
		Mean	Standard deviation	Minimum	Maximum	Mean	Standard deviation	Minimum	Maximum
1	9	6.75	0.46	6.00	7.00	7.00	0.53	6.00	8.00
2	8	6.38	0.74	5.00	7.00	6.38	0.74	5.00	7.00
3	8	5.88	0.64	5.00	7.00	5.88	0.64	5.00	7.00
4	10	7.38	0.52	7.00	8.00	7.38	0.52	7.00	8.00
5	11	6.63	0.92	5.00	8.00	6.63	0.92	5.00	8.00

Table 4.4

Descriptive Statistics for Direct Consensus Method Ratings: Panel A (Afternoon)

Cluster	Number of items	Round one				Round two			
		Mean	Standard deviation	Minimum	Maximum	Mean	Standard deviation	Minimum	Maximum
1	9	6.29	0.49	6.00	7.00	6.29	0.49	6.00	7.00
2	10	6.57	0.53	6.00	7.00	6.57	0.53	6.00	7.00
3	7	4.43	0.53	4.00	5.00	4.43	0.53	4.00	5.00
4	9	5.43	0.53	5.00	6.00	5.43	0.53	5.00	6.00
5	10	6.00	1.00	5.00	7.00	6.43	0.53	6.00	7.00

Table 4.5

Relationship Between Mean Cluster p -value and Panelist Ratings
for Direct Consensus Method: Panel B (Morning)

Cluster	Mean p -value	Percent of items needing to be answered correctly	
		Round one	Round two
1	0.53	0.75	0.78
2	0.70	0.80	0.80
3	0.61	0.73	0.73
4	0.66	0.74	0.74
5	0.55	0.60	0.60
Correlation	--	0.59	0.46

Note. All values were rounded for presentation in table.
Correlations were calculated before rounding took place.

Table 4.6

Relationship Between Mean Cluster p -value and Panelist Ratings
for Direct Consensus Method: Panel A (Afternoon)

Cluster	Mean p -value	Percent of items needing to be answered correctly	
		Round one	Round two
1	0.60	0.70	0.70
2	0.65	0.66	0.66
3	0.59	0.63	0.63
4	0.57	0.60	0.60
5	0.51	0.60	0.64
Correlation	--	0.64	0.30

Note. All values were rounded for presentation in table.
Correlations were calculated before rounding took place.

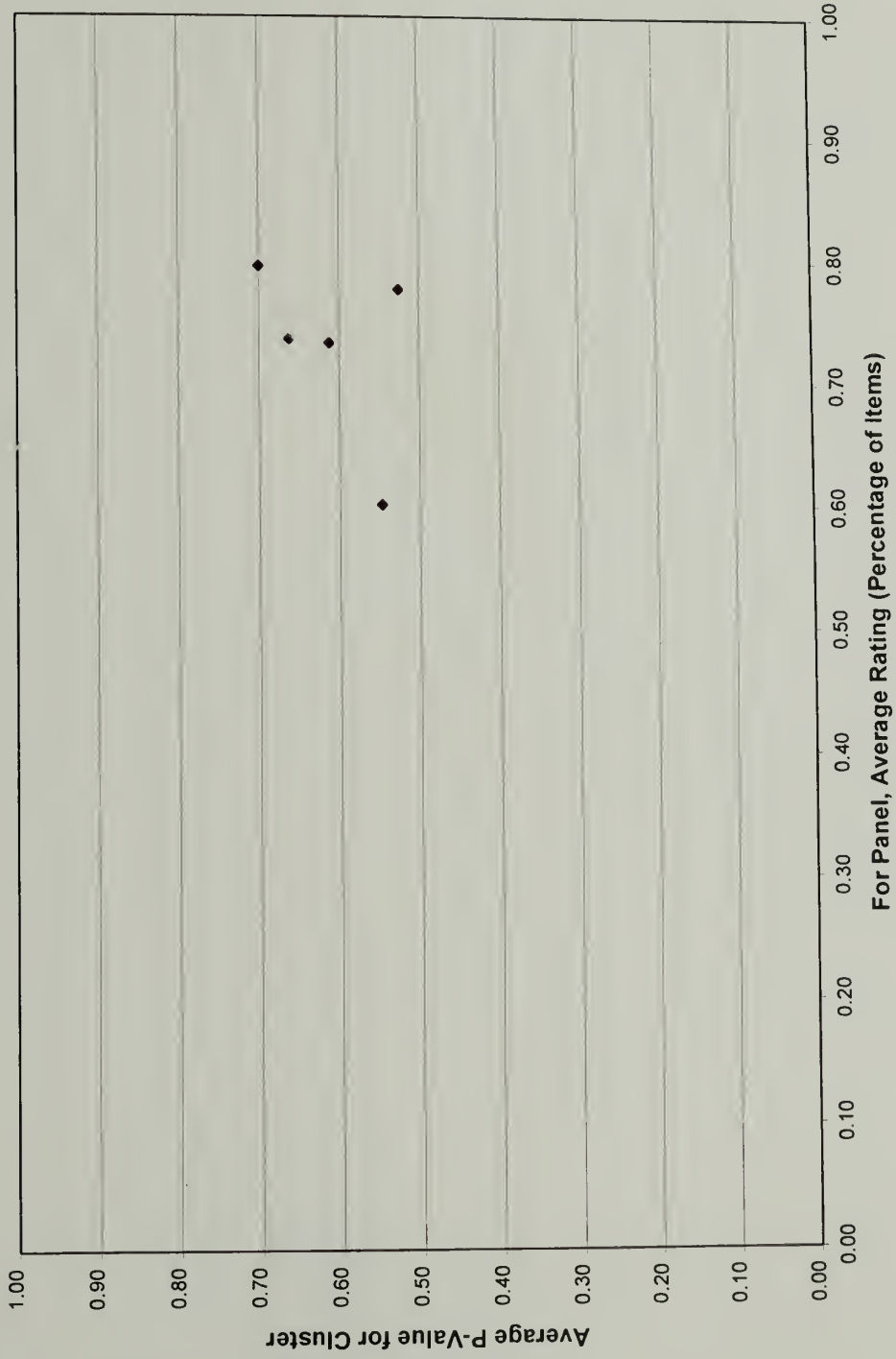


Figure 4.1. Mean panelist cluster rating, in terms of percentage of items, and mean cluster p -value for Panel B (Morning), Round 2, Direct Consensus method.

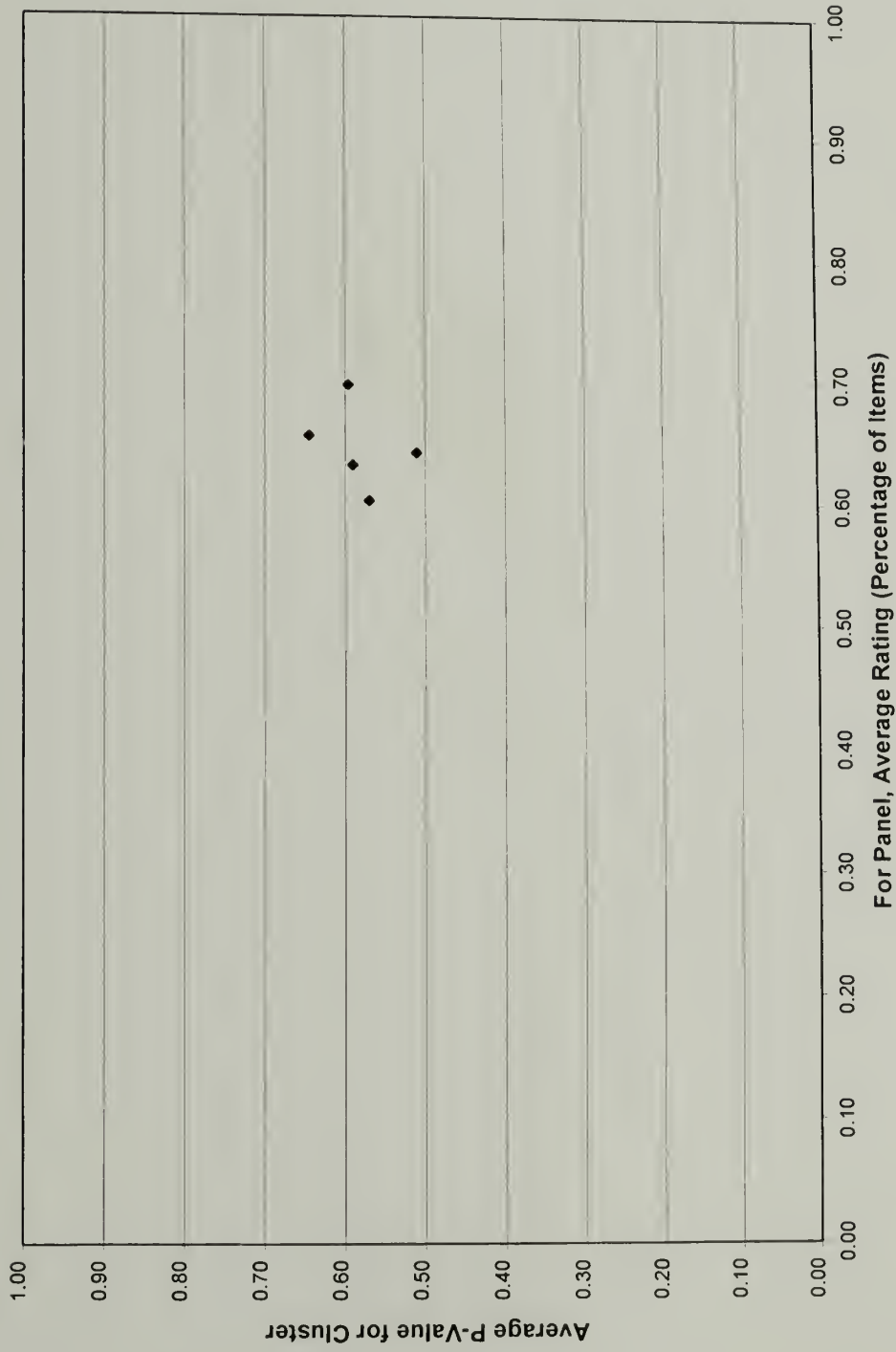


Figure 4.2. Mean panelist cluster rating, in terms of percentage of items, and mean cluster p -value for Panel A (Afternoon), Round 2, Direct Consensus method.

4.2 Item Cluster Method

In contrast to the Direct Consensus method, a cut score is not set during the actual standard-setting session with the Item Cluster method. Instead, panelist ratings can be used in several different ways after the conclusion of the session to determine a cut score. Three types of approaches were used to calculate cut scores in this study; each will be reviewed in turn. Tables 4.7 and 4.8 display the cut scores obtained using each method (for all analyses with the Item Cluster method, results are presented for Panel A first, since they were the first panel to use this method on the day of the study). After the three approaches are described, detailed information on panelist ratings from which the cut scores were calculated is presented.

4.2.1 Boundary Method

In the boundary method, scores for those examinees whose profiles were placed by panelists into one of the borderline rating categories were averaged (see, e.g., Jaeger & Mills, 2001; Plake and Hambleton, 2000, 2001). In this study, ratings for each cluster on the test form were examined separately. All profiles (patterns of right and wrong answers) that were rated as either a 3 (“just below borderline”) or a 4 (“just above borderline”) on the six-point scale were the focus of analysis. The scores, or number of items that the examinees got right in that cluster, were averaged for those examinees whose profiles were rated as a 3 or 4.

The boundary method was implemented in both an individual panelist and aggregated panel manner. That is, in the first approach a boundary-method cut score

Table 4.7

Cut Scores for Panel A (Morning): Item Cluster Method

Panelist	Cut score calculation method				
	Boundary	Equating	Regression		
			Linear	Quadratic	Cubic
A1	28.01	27.61	28.13	28.54	28.57
A2	27.66	28.88	28.40	28.79	28.46
A3	26.40	26.11	26.13	26.16	26.12
A4	27.39	29.00	27.97	27.89	27.62
A5	26.13	27.06	26.63	26.33	27.06
A6	28.10	28.89	27.95	28.85	28.81
A7	27.83	29.01	28.24	28.52	28.55
A8	28.08	28.63	28.63	29.13	29.43
Mean for panel	27.45	28.15	27.76	28.02	28.08
SD	0.77	0.39	0.89	1.16	1.07
By panel ^a	27.42	27.93	27.78	28.06	27.58

^aBy-panel cut scores were calculated by pooling all data for the panel instead of calculating panelist cut scores and averaging them.

Table 4.8

Cut Scores for Panel B (Afternoon): Item Cluster Method

Panelist	Cut score calculation method				
	Boundary	Equating	Regression		
			Linear	Quadratic	Cubic
B1	25.72	24.93	25.05	24.70	24.32
B2	27.05	26.00	26.22	26.41	26.59
B3	25.50	25.38	25.33	25.01	25.02
B4	26.71	24.95	25.46	25.62	25.53
B5	27.18	26.55	26.48	26.69	26.88
B6	26.83	26.55	26.17	26.65	25.99
B7	26.21	28.41	26.17	26.83	27.24
B8	28.92	28.86	27.38	28.82	28.75
Mean for panel	26.77	26.45	26.03	26.34	26.29
SD	1.06	1.49	0.74	1.29	1.39
By panel ^a	26.56	26.00	25.77	26.26	26.02

^aBy-panel cut scores were calculated by pooling all data for the panel instead of calculating panelist cut scores and averaging them.

was calculated for each panelist, and these cut scores were then averaged to obtain a panel cut score for each cluster. Those cluster cut scores were then summed to obtain a test-form cut score. In addition, the boundary method was used on the rounded mean panel ratings for each cluster; those cut scores were then summed across clusters to obtain a test form cut score. The results were very similar for both approaches (by panelist and by panel), as shown in Tables 4.7 and 4.8.

The boundary method is simple to implement, and does not raise issues about the properties of the measurement scale since it treats the ratings as ordinal. However, resulting cut scores are based on only a limited sampling of examinee work (Jaeger & Mills, 2001; Plake & Hambleton, 2001). For that reason, cut score calculation methods that utilize ratings for all examinees' profiles were also evaluated, as described in the next two sections.

4.2.2 Regression Method

Several regression analyses were performed on the data in order to investigate the relationship between examinee profile scores and panelist ratings (see, e.g., Jaeger & Mills, 2001; Plake & Hambleton, 2000, 2001). As with the boundary method, cut scores for each cluster were calculated in two ways: (1) by calculating a cut score for each individual panelist and averaging to obtain a panel cut score, and (2) by averaging panelist ratings and then calculating a panel cut score on those mean ratings.

Three types of regression analyses were performed: linear, quadratic, and cubic. In each case, a model was fit whereby examinee profile scores were considered a function of panelist ratings on the six-point scale. Using the resulting equation,

expected profile scores were calculated for panelist ratings of 3 and 4; those expected profile scores were then averaged to obtain a score for a panelist rating of 3.5.²

Resulting cut scores are displayed in Tables 4.7 and 4.8. A comparison of the values obtained using the linear, quadratic, and cubic methods reveals minor differences. Tables 4.9 and 4.10 display *r*-squared values, representing the percentage of variance accounted for, for each approach. Values are given by cluster, using the method in which ratings were averaged across panelists before the model was fit. As would be expected, fit improved as more parameters were added to the model.

In contrast to the boundary method, in these regression analyses all panelist ratings were used to determine cut scores. However, use of this method assumes that the scale is of an interval nature. In several studies (Jaeger & Mills, 2001; Plake and Hambleton, 2000, 2001), the scale has been adjusted to reflect the smaller semantic differences between ratings adjacent to the standard of interest.

In the current study, exploratory analyses were conducted to determine whether changing values on the rating scale from (1, 2, 3, 4, 5, 6) to (1, 2, 2.75, 3.25, 4, 5) would affect the resulting cut scores. The method in which ratings were averaged across panelists before the model was fit was used for these analyses. The results presented in Tables 4.11 and 4.12 indicate that differences found in the cut scores ranged from 0.01

² An alternative approach would be to directly calculate the expected profile score for a panelist rating of 3.5 (see, e.g., Plake & Hambleton, 1998). This approach was used on an exploratory basis to calculate cut scores using average panelist ratings, and minimal differences were found between these cut scores and those calculated using the average of ratings of 3 and 4. Across the three types of regression (linear, quadratic, cubic) for both panels, there were minimal differences. Four of the six cut scores differed by 0.03 or less, and the largest difference was 0.07. For that reason, no further analyses (i.e., by panelist) were conducted using this approach.

Table 4.9

Percentage of Variance Accounted for by Regression Model:Panel A (Morning)

Cluster ^a	Linear	Quadratic	Cubic
1	0.85	0.87	0.87
2	0.87	0.87	0.89
3	0.88	0.88	0.89
4	0.88	0.89	0.89
5	0.84	0.84	0.85
Mean	0.86	0.87	0.88

^aShown here are results from when models were fit by pooling all data for the panel (versus when models were fit for each panelist).

Table 4.10

Percentage of Variance Accounted for by Regression Model:Panel B (Afternoon)

Cluster ^a	Linear	Quadratic	Cubic
1	0.86	0.86	0.87
2	0.87	0.87	0.87
3	0.88	0.90	0.90
4	0.85	0.85	0.85
5	0.91	0.91	0.91
Mean	0.87	0.88	0.88

^aShown here are results from when models were fit by pooling all data for the panel (versus when models were fit for each panelist).

Table 4.11

Difference in Cut Score and Percentage of Variance Accounted for by
Regression Model When Using Rescaled Ratings: Panel A (Morning)

Cluster	Difference in cut score and <i>r</i> -squared values		
	Linear	Quadratic	Cubic
1	-0.01 (0.00)	-0.04 (0.00)	0.38 (0.00)
2	0.02 (-0.04)	-0.07 (-0.04)	0.36 (0.00)
3	0.03 (-0.03)	-0.02 (-0.03)	-0.06 (0.00)
4	-0.02 (0.00)	-0.05 (0.00)	-0.09 (0.00)
5	-0.01 (0.01)	-0.01 (0.01)	-0.18 (0.00)
Total difference (cut score)	0.01	-0.19	0.41
Mean absolute value difference (<i>r</i> -squared value)	0.02	0.02	0.00

Note. For clusters, cut score difference appears first, followed by the *r*-squared value difference in parentheses. For the purpose of these comparisons, models were fit by pooling all data for the panel instead of fitting models for each panelist.

Table 4.12

Difference in Cut Score and Percentage of Variance Accounted for by
Regression Model When Using Rescaled Ratings: Panel B (Afternoon)

Cluster	Difference in cut score and <i>r</i> -squared values		
	Linear	Quadratic	Cubic
1	-0.01 (0.01)	0.14 (0.01)	0.33 (0.00)
2	0.32 (0.00)	0.05 (0.00)	0.10 (0.00)
3	-0.02 (-0.02)	0.04 (-0.02)	-0.10 (0.00)
4	0.02 (-0.01)	0.00 (-0.01)	0.09 (0.00)
5	0.03 (-0.01)	-0.15 (-0.01)	0.07 (0.00)
Total difference (cut score)	0.34	0.08	0.49
Mean absolute value difference (<i>r</i> -squared value)	0.01	0.01	0.00

Note. For clusters, cut score difference appears first, followed by the *r*-squared value difference in parentheses. For the purpose of these comparisons, models were fit by pooling all data for the panel instead of fitting models for each panelist.

to 0.49, with the biggest differences found with the cubic regression analysis. However, the *r*-squared values indicate no improvement in fit using these rescaled values.

Because of the small affect on cut score and lack of improvement in fit, no further analyses (i.e., by panelist) were conducted using rescaled rating values.

4.2.3 Equating Method

An additional method that may be used to calculate cut scores given ratings such as those obtained in the Item Cluster method utilizes a form of equating to investigate the relationship between examinee profile scores and panelist ratings. Cohen, Kane, and Crooks (1999) noted that regression approaches such as those described above, while minimizing the sum of squared deviations of the profile scores from the regression line, also have the unwanted effect of introducing artifacts such as regression to the mean.

For the current study, the equating approach was applied in two ways: (1) by individual panelist, then summed over panelists for the cluster, and then summed over clusters to obtain a test form cut score, and (2) by calculating a cut score using aggregated panel data for each cluster, and then summing over clusters to obtain a test form cut score. For both approaches, the first step was to count the number of examinee profiles that obtained a panelist rating of 3 or lower; then, the percentage of panelist ratings that this represented was calculated. Next, the examinee profile score that would have the same percentage of the profile distribution below it was determined. That score was used as the cut score for that cluster. Results shown in Tables 4.7 and 4.8 reveal that the cut scores obtained by the individual panelist approach are very close

to those obtained by the aggregated panel data approach. In addition, they are very close to those obtained with the boundary method and regression methods.

4.2.4 Detailed Panelist Rating Information

Information about the panelist ratings on which the Item Cluster method cut scores were based is summarized in this section. First, descriptive statistics for each cluster, for each panel, are presented in Tables 4.13 to 4.22. The mean, standard deviation, minimum and maximum panelist rating for each examinee profile are given for each examinee profile within the cluster; ratings are presented for both rounds one and two. Inspection of these tables reveals that for the majority of examinee profile scores, panelists did not differ more than one rating point from each other.

Also of interest are the number of changes made by panelists between rounds one and two, presented in Tables 4.23 and 4.24. Panelists in the morning (Panel A) made more changes than those in the afternoon (Panel B).

A third type of analysis of panelist ratings involves looking at the relationship between panelist ratings and actual examinee performance. For each cluster, correlations were calculated, for each panelist, between two sets of data: (1) the actual cluster score associated with a given examinee profile, and (2) the rating, on the six-point scale, assigned by the panelist. These correlations are presented in Tables 4.25 and 4.26; means across clusters (for each panelist), across panelists (for each cluster), and a grand mean are also included, for both rounds. The correlations appear to reflect a reasonable degree of relationship between panelist ratings and actual examinee performance. The mean correlation for each panel, for both rounds, was 0.95. For

Panel A, the lowest cluster-based correlation for an individual panelist was 0.90 in Round One, and 0.88 in Round Two. For Panel B, the lowest value was 0.88 for both rounds. Maximum values were, for Panel A, 0.98 for both rounds; for Panel B, 0.99 for both rounds.

Scatterplots showing the location of the data points used in the analyses for Round Two are presented in Figures 4.3 to 4.12. It should be noted that many data points (representing individual panelists' information) overlap; however, the scatterplots are useful in their display of the overall placement and range of the examinee profile scores and panelist ratings.

Table 4.13

Descriptive Statistics for Item Cluster Method Ratings: Panel A (Morning), Cluster 1

Examinee ID	Round one				Round two			
	Mean	SD	Minimum	Maximum	Mean	SD	Minimum	Maximum
04	1.00	0.00	1	1	1.00	0.00	1	1
11	1.75	0.46	1	2	1.75	0.46	1	2
12	2.00	0.00	2	2	2.00	0.00	2	2
02	2.25	0.46	2	3	2.25	0.46	2	3
33	2.75	0.46	2	3	2.88	0.64	2	4
45	2.38	0.52	2	3	2.38	0.52	2	3
17	3.25	0.71	2	4	3.38	0.52	3	4
29	3.13	0.64	2	4	3.25	0.46	3	4
37	3.13	0.35	3	4	3.13	0.35	3	4
89	3.13	0.35	3	4	3.00	0.00	3	3
05	3.81	0.75	3	5	3.75	0.71	3	5
09	4.25	0.71	3	5	4.13	0.83	3	5
24	4.25	0.71	3	5	4.38	0.52	4	5
84	4.13	0.83	3	5	3.88	0.83	3	5
03	4.88	0.35	4	5	4.88	0.35	4	5
30	4.63	0.52	4	5	4.63	0.52	4	5
47	4.69	0.70	4	6	4.50	0.53	4	5
01	5.38	0.52	5	6	5.25	0.46	5	6
131	5.63	0.52	5	6	5.63	0.52	5	6
153	5.56	0.50	5	6	5.75	0.46	5	6

Table 4.14

Descriptive Statistics for Item Cluster Method Ratings: Panel A (Morning), Cluster 2

Examinee ID	Round one				Round two			
	Mean	SD	Minimum	Maximum	Mean	SD	Minimum	Maximum
11	1.38	0.52	1	2	1.38	0.52	1	2
15	1.63	0.52	1	2	1.63	0.52	1	2
51	1.88	0.64	1	3	1.75	0.46	1	2
140	2.06	0.56	1	3	1.88	0.35	1	2
26	2.69	0.46	2	3	2.75	0.46	2	3
45	2.59	0.73	2	4	2.50	0.53	2	3
56	2.75	0.46	2	3	2.75	0.46	2	3
108	3.00	0.00	3	3	3.00	0.00	3	3
04	3.28	0.45	3	4	3.25	0.46	3	4
31	3.63	0.52	3	4	3.63	0.52	3	4
43	3.59	0.50	3	4	3.63	0.52	3	4
57	3.59	0.50	3	4	3.50	0.53	3	4
02	4.06	0.18	4	4.5	4.00	0.00	4	4
14	4.19	0.37	4	5	4.13	0.35	4	5
18	4.38	0.52	4	5	4.38	0.52	4	5
01	5.25	0.46	5	6	5.25	0.46	5	6
20	5.25	0.46	5	6	5.25	0.46	5	6
79	5.25	0.46	5	6	5.25	0.46	5	6

Table 4.15

Descriptive Statistics for Item Cluster Method Ratings: Panel A (Morning), Cluster 3

Examinee ID	Round one				Round two			
	Mean	SD	Minimum	Maximum	Mean	SD	Minimum	Maximum
53	1.25	0.46	1	2	1.25	0.46	1	2
08	1.75	0.46	1	2	1.75	0.46	1	2
51	1.81	0.53	1	2.5	1.81	0.53	1	2.5
115	1.75	0.46	1	2	1.75	0.46	1	2
11	2.66	0.55	2	3.25	2.63	0.52	2	3
15	2.78	0.49	2	3.25	2.75	0.46	2	3
50	2.84	0.35	2	3	2.88	0.35	2	3
83	2.75	0.46	2	3	2.75	0.46	2	3
02	3.59	0.50	3	4	3.63	0.52	3	4
05	3.50	0.53	3	4	3.63	0.52	3	4
07	3.41	0.57	3	4.25	3.50	0.53	3	4
18	4.00	0.13	3.75	4.25	4.00	0.00	4	4
01	4.13	0.35	4	5	4.13	0.35	4	5
03	4.63	0.52	4	5	4.63	0.52	4	5
67	4.50	0.53	4	5	4.50	0.53	4	5
06	5.34	0.44	5	6	5.19	0.37	5	6
10	5.38	0.52	5	6	5.38	0.52	5	6
87	5.16	0.35	5	6	5.13	0.35	5	6

Table 4.16

Descriptive Statistics for Item Cluster Method Ratings: Panel A (Morning), Cluster 4

Examinee ID	Round one				Round two			
	Mean	SD	Minimum	Maximum	Mean	SD	Minimum	Maximum
15	1.19	0.37	1	2	1.00	0.00	1	1
08	1.38	0.52	1	2	1.38	0.52	1	2
40	1.25	0.46	1	2	1.25	0.46	1	2
65	2.00	0.00	2	2	2.00	0.00	2	2
04	2.72	0.45	2	3	2.75	0.46	2	3
43	2.72	0.45	2	3	2.88	0.35	2	3
52	2.47	0.51	2	3	2.63	0.52	2	3
72	2.72	0.45	2	3	2.75	0.46	2	3
126	2.59	0.50	2	3	2.63	0.52	2	3
09	3.34	0.48	3	4	3.38	0.52	3	4
17	3.38	0.52	3	4	3.38	0.52	3	4
47	3.25	0.46	3	4	3.25	0.46	3	4
54	3.63	0.52	3	4	3.63	0.52	3	4
92	3.25	0.46	3	4	3.25	0.46	3	4
01	4.22	0.41	4	5	4.25	0.46	4	5
24	4.28	0.45	4	5	4.25	0.46	4	5
60	4.00	0.53	3	5	3.88	0.64	3	5
14	4.88	0.35	4	5	4.88	0.35	4	5
98	4.91	0.38	4	5.25	4.88	0.35	4	5
32	5.81	0.37	5	6	6.00	0.00	6	6

Table 4.17

Descriptive Statistics for Item Cluster Method Ratings: Panel A (Morning), Cluster 5

Examinee ID	Round one				Round two			
	Mean	SD	Minimum	Maximum	Mean	SD	Minimum	Maximum
27	1.00	0.00	1	1	1.00	0.00	1	1
17	2.00	0.76	1	3	2.13	0.64	1	3
14	2.25	1.04	1	4	2.38	0.74	1	3
68	2.31	0.80	1	3.5	2.38	0.74	1	3
04	2.72	0.65	2	3.75	2.88	0.35	2	3
75	3.00	0.53	2	4	3.13	0.35	3	4
87	3.00	0.76	2	4	3.25	0.46	3	4
06	3.25	0.46	3	4	3.38	0.52	3	4
32	3.25	0.46	3	4	3.38	0.52	3	4
44	3.63	0.92	2	5	3.88	0.83	3	5
90	3.88	0.64	3	5	4.00	0.76	3	5
07	4.13	0.83	3	5	4.75	0.46	4	5
28	4.25	0.89	3	6	4.38	0.92	3	6
96	4.25	0.71	3	5	4.38	0.74	3	5
09	4.88	0.35	4	5	5.00	0.53	4	6
02	4.88	0.35	4	5	5.00	0.53	4	6
38	5.00	0.53	4	6	5.13	0.64	4	6
102	5.75	0.46	5	6	5.75	0.46	5	6
01	6.00	0.00	6	6	6.00	0.00	6	6
329	6.00	0.00	6	6	6.00	0.00	6	6

Table 4.18

Descriptive Statistics for Item Cluster Method Ratings: Panel B (Afternoon), Cluster 1

Examinee ID	Round one				Round two			
	Mean	SD	Minimum	Maximum	Mean	SD	Minimum	Maximum
04	1.00	0.00	1	1	1.00	0.00	1	1
11	1.50	0.53	1	2	1.50	0.53	1	2
12	1.63	0.52	1	2	1.63	0.52	1	2
02	2.38	0.52	2	3	2.38	0.52	2	3
33	2.25	0.46	2	3	2.25	0.46	2	3
45	2.25	0.46	2	3	2.25	0.46	2	3
17	3.38	0.52	3	4	3.38	0.52	3	4
29	3.63	0.74	3	5	3.63	0.74	3	5
37	3.63	0.52	3	4	3.63	0.52	3	4
89	3.63	0.52	3	4	3.63	0.52	3	4
05	4.13	0.83	3	5	4.13	0.83	3	5
09	4.38	0.52	4	5	4.38	0.52	4	5
24	4.25	0.89	3	5	4.25	0.89	3	5
84	4.25	0.89	3	5	4.25	0.89	3	5
03	4.88	0.64	4	6	4.75	0.46	4	5
30	5.00	0.53	4	6	4.88	0.35	4	5
47	5.00	0.00	5	5	5.00	0.00	5	5
01	5.75	0.46	5	6	5.75	0.46	5	6
131	5.50	0.53	5	6	5.50	0.53	5	6
153	6.00	0.00	6	6	6.00	0.00	6	6

Table 4.19

Descriptive Statistics for Item Cluster Method Ratings: Panel B (Afternoon), Cluster 2

Examinee ID	Round one				Round two			
	Mean	SD	Minimum	Maximum	Mean	SD	Minimum	Maximum
736	1.00	0.00	1	1	1.00	0.00	1	1
11	1.38	0.52	1	2	1.38	0.52	1	2
36	2.13	0.35	2	3	2.13	0.35	2	3
76	2.13	0.35	2	3	2.13	0.35	2	3
244	2.00	0.53	1	3	2.00	0.53	1	3
04	2.63	0.52	2	3	2.63	0.52	2	3
22	2.00	0.76	1	3	2.00	0.76	1	3
94	2.63	0.52	2	3	2.63	0.52	2	3
135	2.50	0.53	2	3	2.50	0.53	2	3
06	3.38	0.52	3	4	3.38	0.52	3	4
19	3.38	0.52	3	4	3.38	0.52	3	4
27	3.38	0.52	3	4	3.38	0.52	3	4
28	4.00	0.00	4	4	4.00	0.00	4	4
12	4.38	0.52	4	5	4.38	0.52	4	5
14	4.38	0.52	4	5	4.38	0.52	4	5
63	5.00	0.00	5	5	5.00	0.00	5	5
01	5.50	0.53	5	6	5.50	0.53	5	6
58	5.13	0.35	5	6	5.13	0.35	5	6
07	6.00	0.00	6	6	6.00	0.00	6	6

Table 4.20

Descriptive Statistics for Item Cluster Method Ratings: Panel B (Afternoon), Cluster 3

Examinee ID	Round one				Round two			
	Mean	SD	Minimum	Maximum	Mean	SD	Minimum	Maximum
54	1.00	0.00	1	1	1.00	0.00	1	1
12	1.63	0.52	1	2	1.63	0.52	1	2
40	1.75	0.46	1	2	1.75	0.46	1	2
07	2.50	0.53	2	3	2.50	0.53	2	3
08	2.38	0.52	2	3	2.38	0.52	2	3
49	2.50	0.53	2	3	2.50	0.53	2	3
01	3.13	0.35	3	4	3.13	0.35	3	4
02	3.25	0.46	3	4	3.25	0.46	3	4
30	3.50	0.53	3	4	3.50	0.53	3	4
59	3.38	0.52	3	4	3.38	0.52	3	4
05	4.13	0.35	4	5	4.13	0.35	4	5
17	4.75	0.46	4	5	4.75	0.46	4	5
36	4.50	0.53	4	5	4.50	0.53	4	5
37	4.25	0.46	4	5	4.25	0.46	4	5
03	5.25	0.46	5	6	5.25	0.46	5	6
13	5.25	0.46	5	6	5.25	0.46	5	6
21	5.88	0.35	5	6	5.88	0.35	5	6

Table 4.21

Descriptive Statistics for Item Cluster Method Ratings: Panel B (Afternoon), Cluster 4

Examinee ID	Round one				Round two			
	Mean	SD	Minimum	Maximum	Mean	SD	Minimum	Maximum
04	1.13	0.35	1	2	1.13	0.35	1	2
08	1.50	0.53	1	2	1.50	0.53	1	2
37	1.75	0.46	1	2	1.75	0.46	1	2
11	1.88	0.64	1	3	1.88	0.64	1	3
64	2.38	0.52	2	3	2.38	0.52	2	3
72	2.50	0.53	2	3	2.50	0.53	2	3
07	3.38	0.74	2	4	3.50	0.53	3	4
14	3.63	1.06	2	5	3.75	0.89	3	5
36	3.63	0.74	3	5	3.63	0.74	3	5
58	3.38	0.52	3	4	3.38	0.52	3	4
06	4.00	0.53	3	5	4.00	0.53	3	5
24	4.00	0.76	3	5	4.00	0.76	3	5
70	3.88	0.35	3	4	3.88	0.35	3	4
131	4.00	0.76	3	5	4.00	0.76	3	5
01	4.75	0.46	4	5	4.75	0.46	4	5
29	4.88	0.35	4	5	5.00	0.00	5	5
92	4.88	0.35	4	5	4.88	0.35	4	5
10	5.50	0.53	5	6	5.50	0.53	5	6
47	5.50	0.53	5	6	5.50	0.53	5	6
78	5.88	0.35	5	6	5.88	0.35	5	6

Table 4.22

Descriptive Statistics for Item Cluster Method Ratings: Panel B (Afternoon), Cluster 5

Examinee ID	Round one				Round two			
	Mean	SD	Minimum	Maximum	Mean	SD	Minimum	Maximum
27	1.13	0.35	1	2	1.13	0.35	1	2
14	1.13	0.35	1	2	1.13	0.35	1	2
17	1.63	0.52	1	2	1.63	0.52	1	2
04	1.63	0.52	1	2	1.63	0.52	1	2
28	1.75	0.46	1	2	1.75	0.46	1	2
75	1.50	0.53	1	2	1.50	0.53	1	2
07	2.00	0.53	1	3	2.00	0.53	1	3
44	2.50	0.53	2	3	2.50	0.53	2	3
09	2.63	0.52	2	3	2.63	0.52	2	3
87	3.50	0.53	3	4	3.50	0.53	3	4
01	3.88	0.64	3	5	3.88	0.64	3	5
68	4.25	0.46	4	5	4.25	0.46	4	5
06	4.63	0.52	4	5	4.63	0.52	4	5
32	4.50	0.53	4	5	4.50	0.53	4	5
96	4.25	0.71	3	5	4.25	0.71	3	5
329	4.75	0.46	4	5	4.75	0.46	4	5
02	4.75	0.46	4	5	4.75	0.46	4	5
90	5.63	0.52	5	6	5.63	0.52	5	6
38	5.88	0.35	5	6	5.88	0.35	5	6
102	6.00	0.00	6	6	6.00	0.00	6	6

Table 4.23

Changes in Ratings by Panelist Across Rounds forItem Cluster Method: Panel A (Morning)

Panelist	Total number of changes	Mean change	Mean absolute change
A1	5	0.03	0.05
A2	8	-0.08	0.08
A3	11	0.02	0.08
A4	8	-0.02	0.03
A5	25	0.00	0.10
A6	16	-0.13	0.17
A7	8	-0.01	0.09
A8	0	0.00	0.00
Panel mean	10	-0.02	0.07

Note. Table includes information on number of changes made, summed across clusters, from round one to round two.

Table 4.24

Changes in Ratings by Panelist Across Rounds for
Item Cluster Method: Panel B (Afternoon)

Panelist	Total number of changes	Mean change	Mean absolute change
B1	0	0.00	0.00
B2	1	0.00	0.01
B3	0	0.00	0.00
B4	0	0.00	0.00
B5	3	0.00	0.03
B6	0	0.00	0.00
B7	0	0.00	0.00
B8	1	0.00	0.01
Panel mean	1	0.00	0.01

Note. Table includes information on number of changes made, summed across clusters, from round one to round two.

Table 4.25

Correlations Between Examinee Profile Scores and Panelist Ratings for Item Cluster Method: Panel A (Morning)

Panelist	Round one					Round two						
	Cluster					Cluster						
	1	2	3	4	5	Mean	1	2	3	4	5	Mean
A1	0.94	0.93	0.97	0.96	0.98	0.96	0.93	0.96	0.97	0.96	0.98	0.96
A2	0.95	0.95	0.95	0.92	0.97	0.95	0.95	0.95	0.95	0.92	0.94	0.94
A3	0.90	0.97	0.95	0.95	0.96	0.95	0.88	0.97	0.95	0.98	0.97	0.95
A4	0.97	0.95	0.96	0.95	0.92	0.95	0.97	0.95	0.96	0.96	0.92	0.95
A5	0.94	0.96	0.95	0.98	0.93	0.95	0.96	0.96	0.95	0.96	0.93	0.95
A6	0.91	0.92	0.93	0.92	0.96	0.93	0.93	0.92	0.95	0.92	0.96	0.94
A7	0.96	0.95	0.98	0.97	0.93	0.96	0.94	0.95	0.98	0.95	0.95	0.95
A8	0.94	0.97	0.95	0.97	0.98	0.96	0.94	0.97	0.95	0.97	0.98	0.96
Mean	0.94	0.95	0.96	0.95	0.95	0.95	0.94	0.95	0.96	0.95	0.95	0.95

Table 4.26

Correlations Between Examinee Profile Scores and Panelist Ratings for Item Cluster Method: Panel B (Afternoon)

Panelist	Round one					Round two						
	Cluster					Cluster						
	1	2	3	4	5	Mean	1	2	3	4	5	Mean
B1	0.96	0.98	0.94	0.96	0.97	0.96	0.96	0.98	0.94	0.96	0.97	0.96
B2	0.94	0.92	0.95	0.94	0.97	0.94	0.94	0.92	0.95	0.94	0.97	0.94
B3	0.96	0.96	0.96	0.90	0.96	0.95	0.96	0.96	0.96	0.90	0.96	0.95
B4	0.96	0.95	0.95	0.88	0.96	0.94	0.96	0.95	0.95	0.88	0.96	0.94
B5	0.97	0.92	0.95	0.95	0.96	0.95	0.97	0.92	0.95	0.95	0.96	0.95
B6	0.98	0.95	0.97	0.95	0.99	0.97	0.98	0.95	0.97	0.95	0.99	0.97
B7	0.95	0.95	0.94	0.96	0.96	0.95	0.95	0.95	0.94	0.96	0.96	0.95
B8	0.94	0.93	0.95	0.94	0.97	0.94	0.94	0.93	0.95	0.95	0.97	0.95
Mean	0.96	0.94	0.95	0.94	0.97	0.95	0.96	0.94	0.95	0.94	0.97	0.95

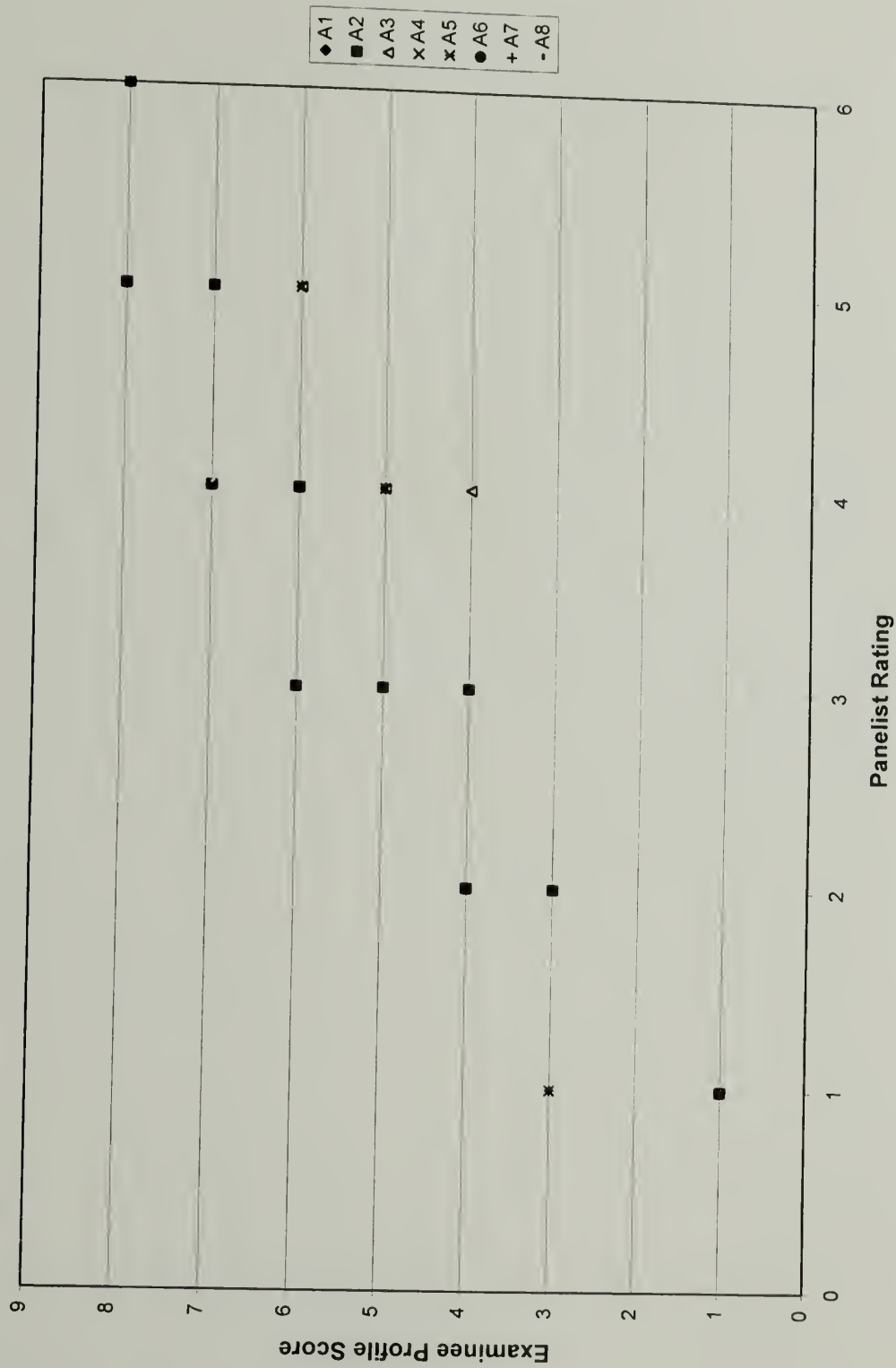


Figure 4.3. Panelist ratings and examinee profile scores for Panel A (Morning), Cluster 1, Round 2, Item Cluster method.

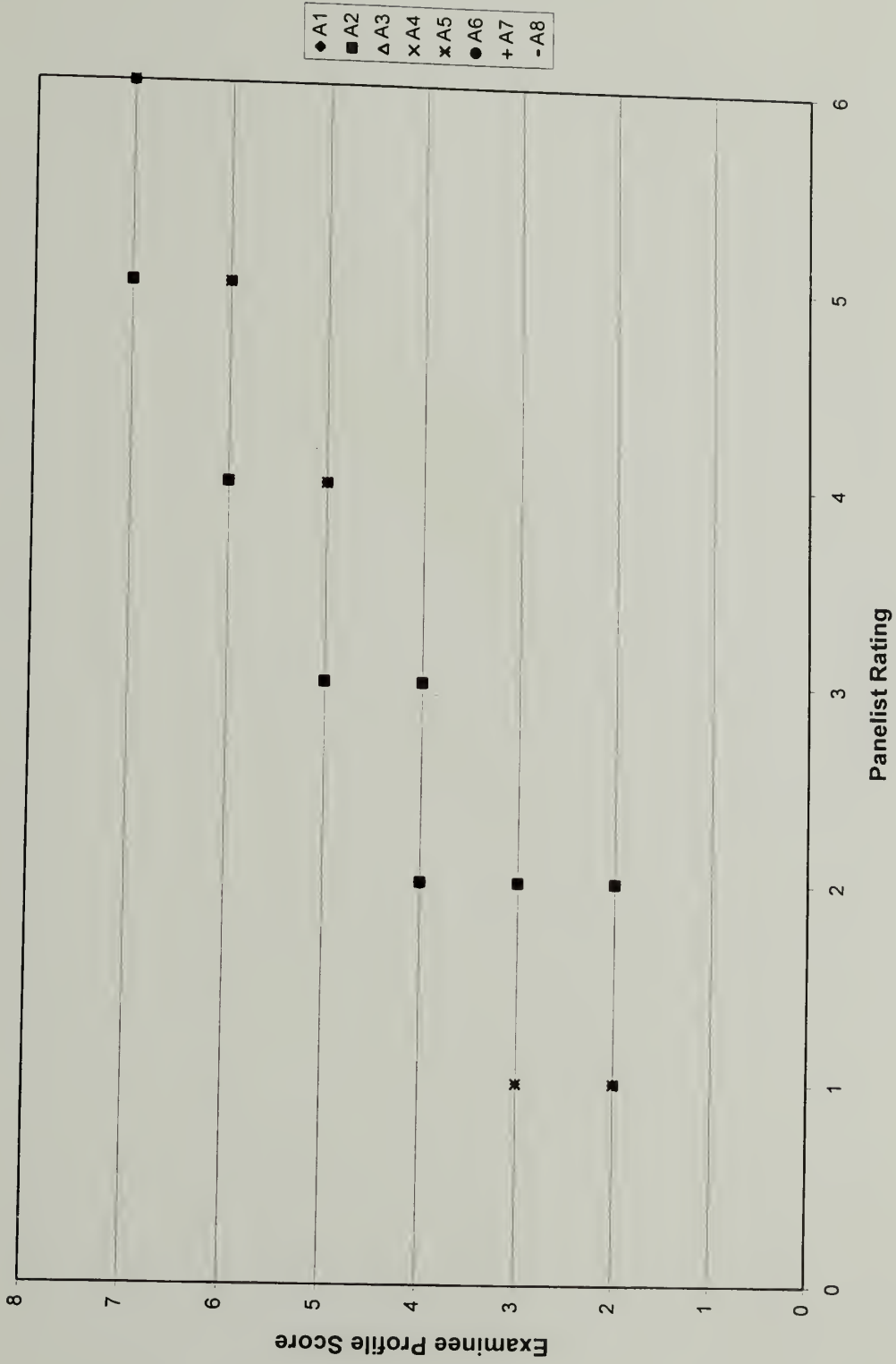


Figure 4.4. Panelist ratings and examinee profile scores for Panel A (Morning), Cluster 2, Round 2, Item Cluster method.



Figure 4.5. Panelist ratings and examinee profile scores for Panel A (Morning), Cluster 3, Round 2, Item Cluster method.

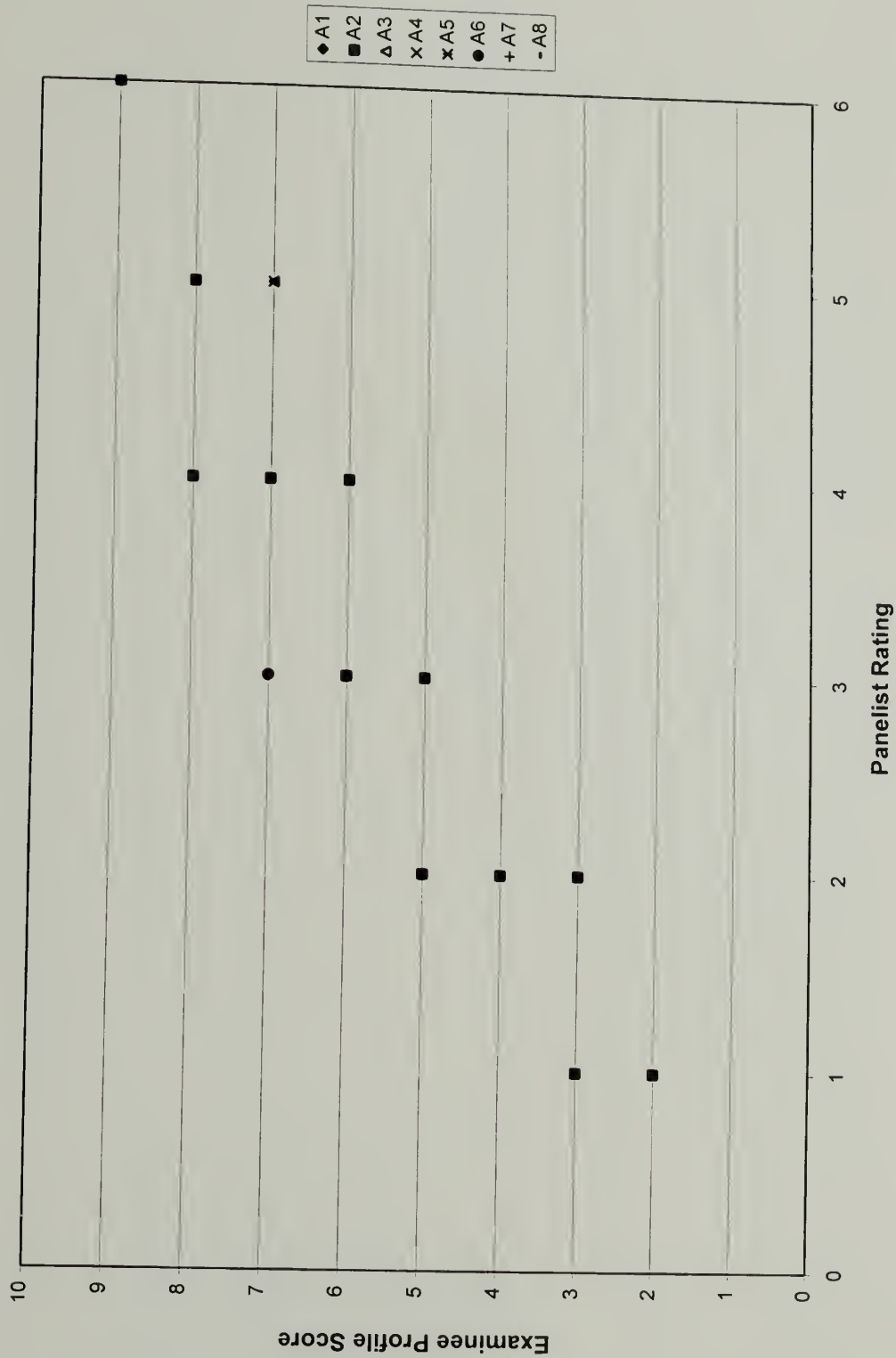


Figure 4.6. Panelist ratings and examinee profile scores for Panel A (Morning), Cluster 4, Round 2, Item Cluster method.

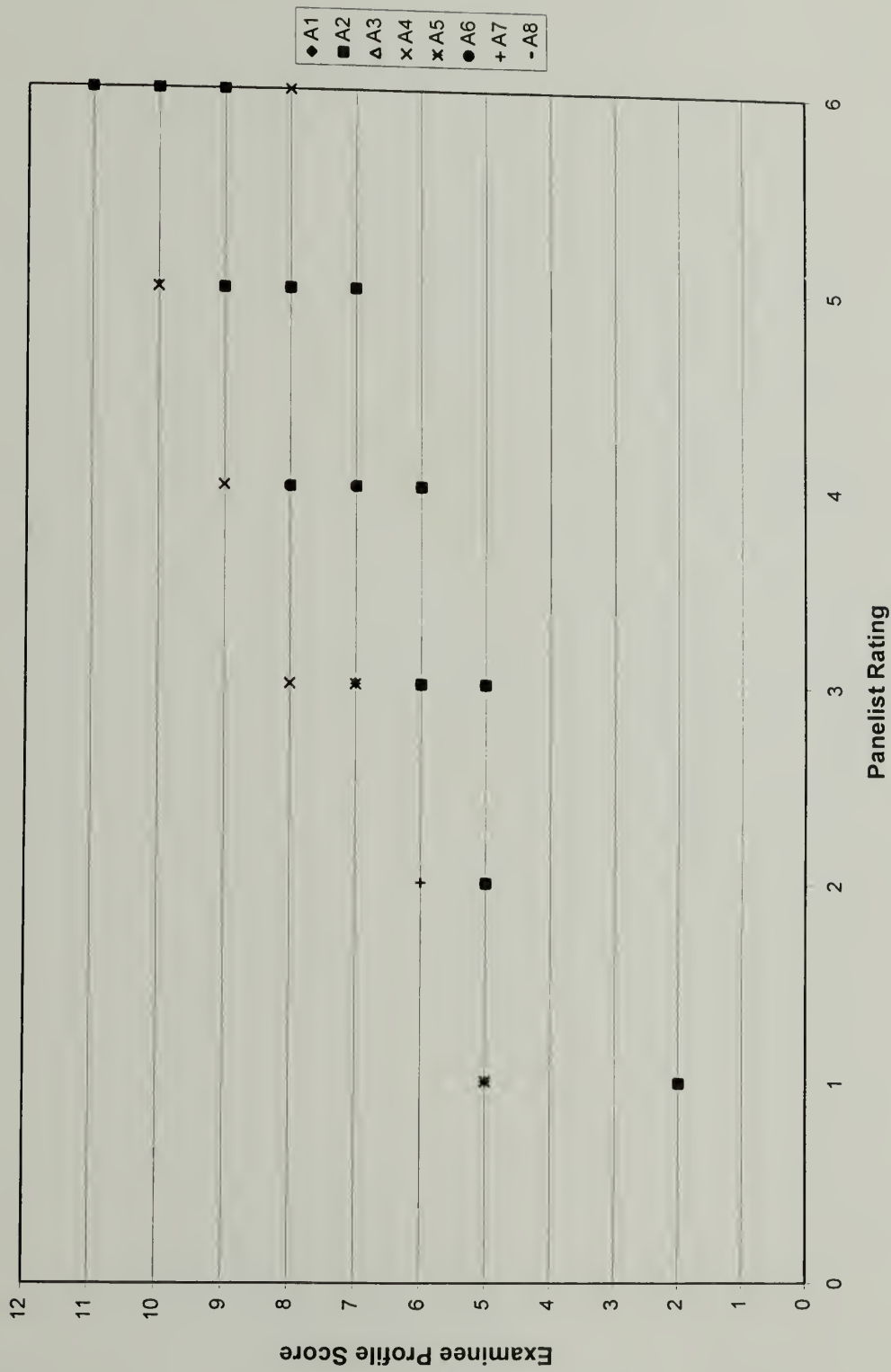


Figure 4.7. Panelist ratings and examinee profile scores for Panel A (Morning), Cluster 5, Round 2, Item Cluster method.



Figure 4.8. Panelist ratings and examinee profile scores for Panel B (Afternoon), Cluster 1, Round 2, Item Cluster method.

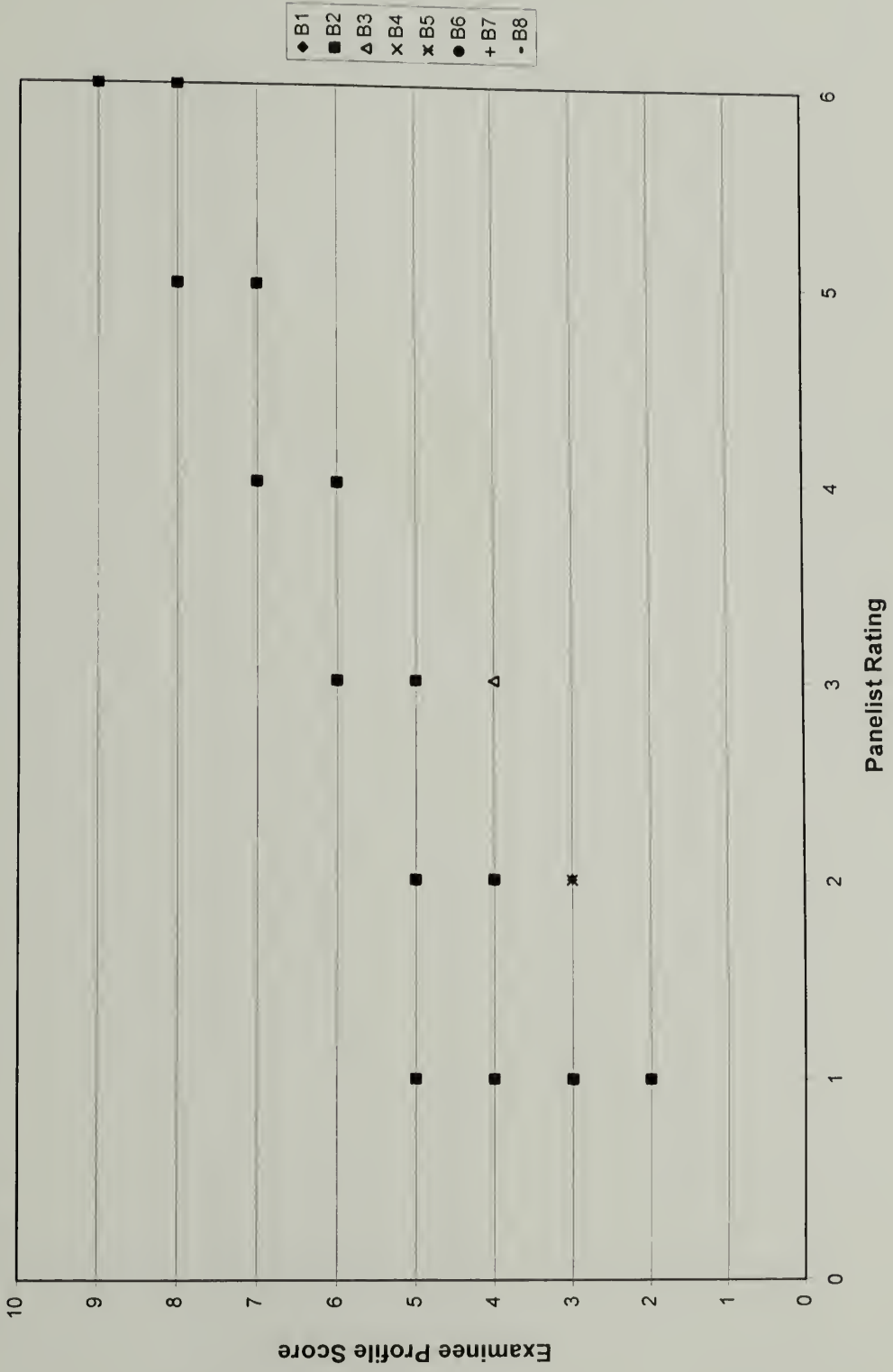


Figure 4.9. Panelist ratings and examinee profile scores for Panel B (Afternoon), Cluster 2, Round 2, Item Cluster method.

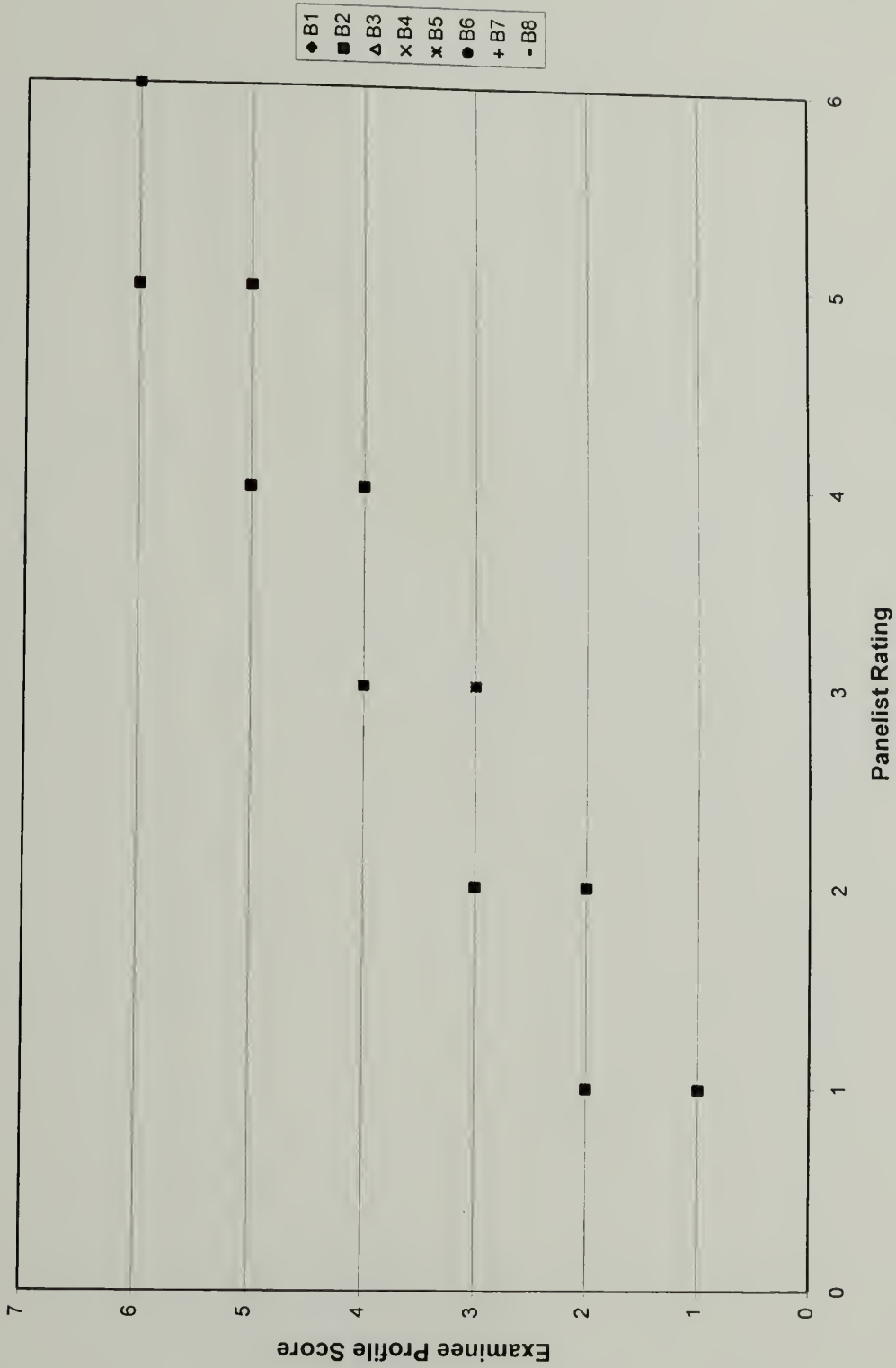


Figure 4.10. Panelist ratings and examinee profile scores for Panel B (Afternoon), Cluster 3, Round 2, Item Cluster method.

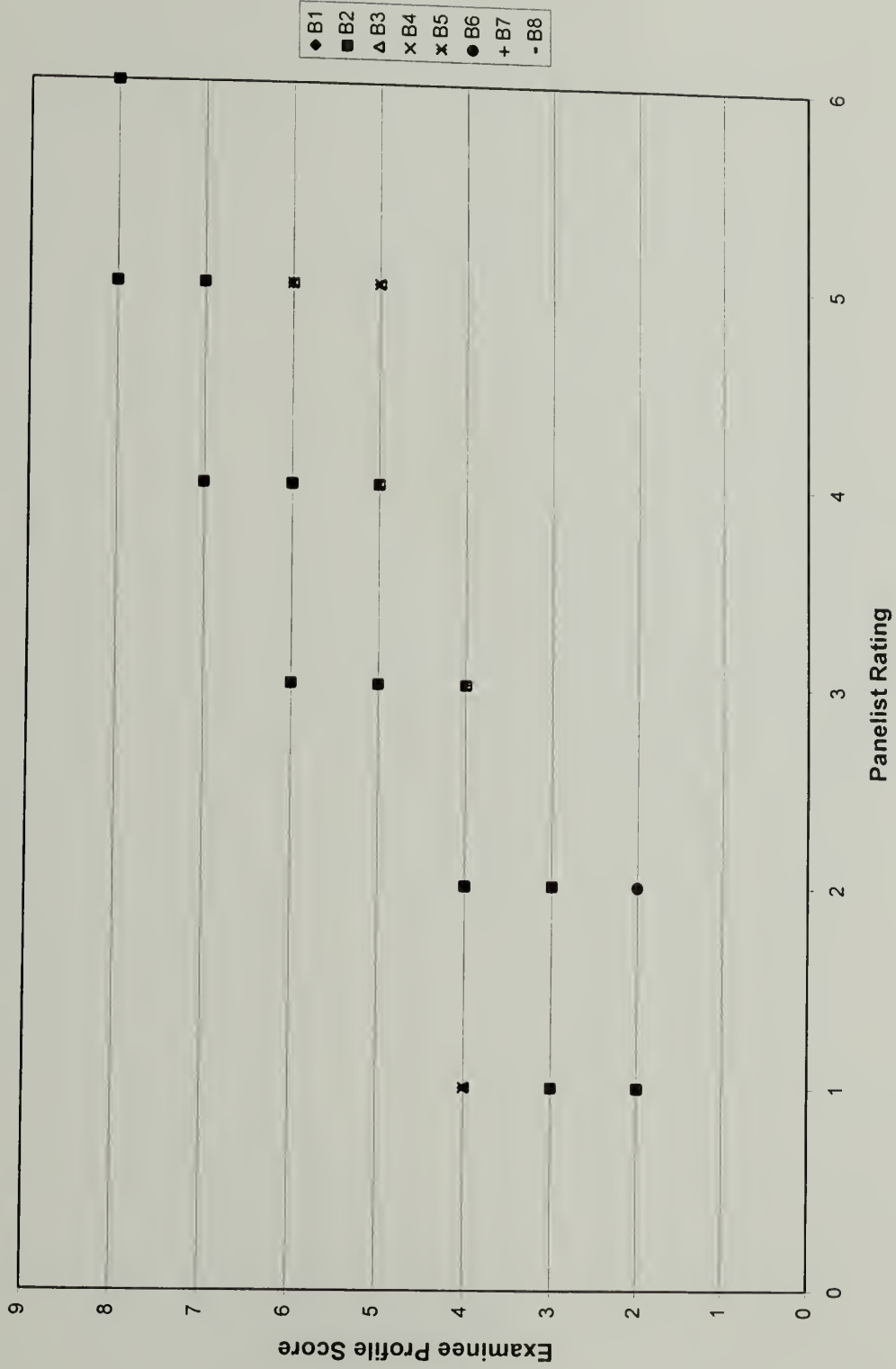


Figure 4.11. Panelist ratings and examinee profile scores for Panel B (Afternoon), Cluster 4, Round 2, Item Cluster method.

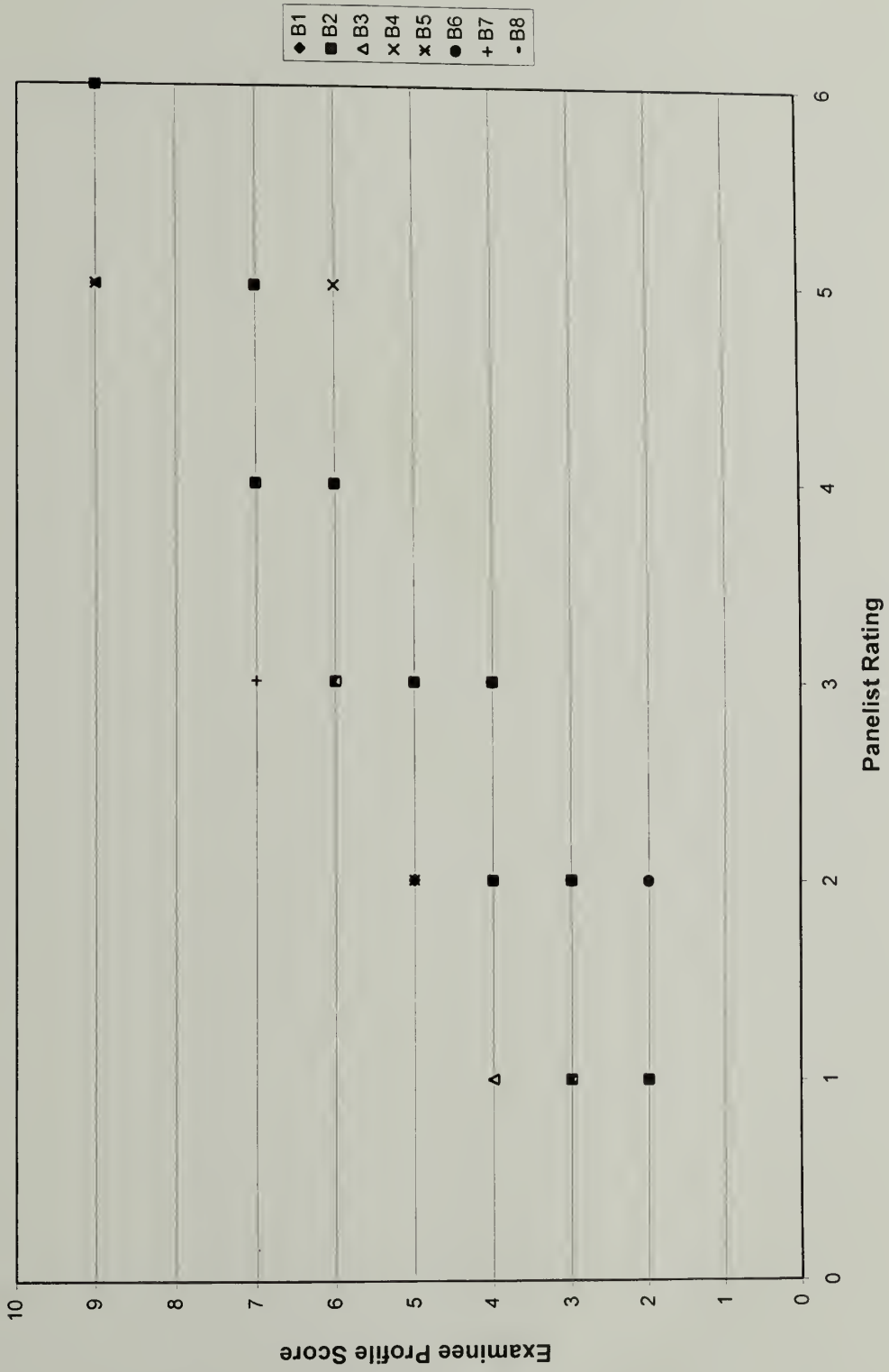


Figure 4.12. Panelist ratings and examinee profile scores for Panel B (Afternoon), Cluster 5, Round 2, Item Cluster method.

4.3 Summary Information

This section contains summary information for both methods. First, the two methods are compared directly across sessions. Timing information is presented next, followed by a summary of responses to the evaluation survey.

4.3.1 Comparison of Cut Scores and Their Impact

In previous sections, the cut scores obtained by the two different standard-setting methods were presented. Within this section, these cut scores are compared directly across methods and sessions.

Table 4.27 summarizes the cut scores obtained using both methods. The percent-correct, in terms of items, that the cut score represents is also given. It is important to note that Test Form 1, used in the morning session, contained 46 items, while Test Form 2, used in the afternoon session, contained 45 items (its OOAF contained one fewer item). The Item Cluster method cut scores are those obtained by calculating separate panelist cut scores and then averaging them (as opposed to the approach whereby all panel data was pooled before calculation of the cut scores). Only cubic regression results are provided for that type of analysis, since that model accounted for more of the variance.

The two methods obviously produced noticeably different cut scores. The Direct Consensus method yielded cut scores that were in all cases higher than Item Cluster method cut scores, regardless of the calculation method of the latter chosen for comparison.

It is also helpful to review evidence of the reasonableness of the cut scores. For this purpose, the impact of these cut scores on the passing rate was also estimated; these values are presented in Table 4.27 as well. The process for estimating the pass rates was as follows. First, the expected mean score on each of the two test forms was calculated by summing the p -values of the items on that form (with p -value representing the proportion of examinees who got each item correct), as shown in formula 1.

$$\bar{X} = \sum_{i=1}^n p_i \quad [1]$$

Next, the standard deviation for each test form score distribution was estimated using the following formula (Lord & Novick, 1968).

$$\sigma_X = \sum_{i=1}^n S_i r_{ix} \quad [2]$$

where S_i is the item standard deviation, or $(p)(1-p)$, and r_{ix} is the item/test score correlation, or item discrimination index. The estimated mean and standard deviation for Test Form 1 were 27.87 and 9.86, respectively; for Test Form 2, these values were 26.17 and 9.72.

If we are willing to make a normality assumption regarding the distribution of examinee scores, we are then able to obtain the z -score for a given cut score using the estimated mean test score and standard deviation. The area under the normal curve to the right of the z -score for a given cut score can then be viewed as the percentage of the examinees who would obtain scores greater than the cut score, thus passing the examination. It should be noted that cut scores were not rounded for the purposes of estimating the pass rates. The rationale for this approach was that number-right scores

would in practice be transformed into scaled scores before rounding was undertaken; thus not rounding allows us to better estimate the eventual impact of different cut scores. The estimated pass rates are, as shown in Table 4.27, clearly higher with the Item Cluster method than with the Direct Consensus method.

Table 4.27

Comparison of Cut Scores Across Methods

Session	Method	Cut score	Percent correct (items)	Estimated percent passing (examinees)
Panel A				
Morning	ICM-Boundary	27.45	60%	52%
	ICM-Regression	28.08	61%	49%
	ICM-Equating	28.15	61%	49%
Afternoon	DCM	30	67%	35%
Panel B				
Afternoon	ICM-Boundary	26.77	59%	48%
	ICM-Regression	26.29	58%	50%
	ICM-Equating	26.45	59%	49%
Morning	DCM	34	74%	27%

Note. ICM = Item Cluster method, DCM = Direct Consensus method. ICM cut scores are those obtained by calculating separate panelist cut scores and then averaging them; cubic regression results are those provided for ICM-regression. Each test form contained 35 multiple-choice questions; however, test form 1 (used in the morning session) contained 11 OOAF items, and test form 2 (used in the afternoon session) contained 10 OOAF items.

4.3.2 Timing Information

An important consideration in the conduct of standard setting is how long it takes to set a standard using a given method. Standard setting panelists in many professions have hectic schedules and are both expensive and difficult to recruit. Therefore, standard setting methods are needed that are psychometrically defensible, but minimize the amount of time needed from expert panelists.

A summary of the time needed for each of the two methods for the morning and afternoon sessions is presented in Table 4.28. As expected due to its less detail-oriented nature, the Direct Consensus method took less time than the Item Cluster method. In the morning session, the Direct Consensus method took only 60% of the time that the Item Cluster method did; in the afternoon, it took 59%.

Both methods took less time in the afternoon, presumably due to the panelists having become comfortable with the concept of the minimally competent CPA and gaining familiarity with the test materials in the morning session. In Tables 4.29 and 4.30, more detailed information about the amount of time needed for each component of the methods is presented.

Table 4.28

Summary of Timing Information

Method	Session	
	Morning	Afternoon
Item Cluster	3 hours, 57 minutes	3 hours, 24 minutes
Direct Consensus	2 hours, 23 minutes	2 hours, 1 minute

Table 4.29

Timing Information: Direct Consensus Method

Component	Morning (Panel B)	Afternoon (Panel A)
Introduction	10 minutes	6 minutes
Cluster 1		
Individual ratings	12 minutes	20 minutes
Review/discussion/ revise ratings	25 minutes	15 minutes
Clusters 2–5		
Individual ratings	61 minutes	40 minutes
Review/discussion/ revise ratings	20 minutes	30 minutes
Arrive at consensus		
Review/discussion/ revise ratings	15 minutes	10 minutes
Total time	2 hours, 23 minutes	2 hours, 1 minute

Table 4.30

Timing Information: Item Cluster Method

Component	Morning (Panel A)	Afternoon (Panel B)
Training/ practice exercise	50 minutes	33 minutes
Cluster 1		
Individual ratings	29 minutes	31 minutes
Review/discussion/ revise ratings	24 minutes	20 minutes
Cluster 2		
Individual ratings	22 minutes	29 minutes
Review/discussion/ revise ratings	17 minutes	10 minutes
Cluster 3		
Individual ratings	32 minutes (panelists ate lunch during this time also)	20 minutes
Review/discussion/ revise ratings	13 minutes	11 minutes
Cluster 4		
Individual ratings	10 minutes	16 minutes
Review/discussion/ revise ratings	11 minutes	11 minutes
Cluster 5		
Individual ratings	14 minutes	23 minutes
Review/discussion/ revise ratings	15 minutes	N/A (ran out of time)
Total time	3 hours, 57 minutes	3 hours, 24 minutes

4.3.2 Evaluation Survey Results

The evaluation survey for the study contained a mixture of (1) three- and five-option Likert-scale items, (2) items to which panelists could indicate as many responses as applied, and (3) open-ended questions. Frequencies for responses to each option for the first two types of items are presented in Tables 4.31 to 4.33. Table 4.31 contains answers to general questions; Tables 4.32 and 4.33 contains answers to questions regarding the Direct Consensus and Item Cluster methods, respectively. Panelist responses to the open-ended questions were transcribed verbatim; these are presented in Table 4.34. Responses to these questions will be discussed as applicable in Chapter 5.

Table 4.31

Evaluation Survey Results: General Questions

Question number and topic	Rating category	Panel	
		A (ICM first) ^a	B (DCM first)
1. Clarity of description of minimally competent CPA	Very clear	50.0% (4)	62.5% (5)
	Clear	37.5% (3)	37.5% (3)
	Somewhat clear	12.5% (1)	--
	Not clear	--	--
26. Method that would produce defensible passing score for the CPA Exam	Confidence in both	25.0% (2)	75.0% (6)
	Confidence in DCM only	62.5% (5)	25.0% (2)
	Confidence in ICM only	--	--
	No confidence in either	--	--
	No response	12.5% (1)	--
29. Recommendation if one method were to be chosen for setting passing scores on CPA Exam	Both	12.5% (1)	12.5% (1)
	DCM	75.0% (6)	62.5% (5)
	ICM	--	25.0% (2)
	No response	12.5% (1)	--

(continued on next page)

Table 4.31 (continued)

Question number and topic	Rating category	Panel	
		A (ICM first) ^a	B (DCM first)
27. Presence of confusion due to the use of two methods for setting passing scores	Not at all confused	87.5% (7)	100.0% (8)
	Occasionally confused	--	--
	Definitely a problem	--	--
	No response	12.5% (1)	--
28. Did participation in first session influence ratings in second session	Yes	12.5% (1)	--
	No	62.5% (5)	100.0% (8)
	No response	25.0% (2)	--

Note. ICM = Item Cluster method; DCM = Direct Consensus method.

^aEight panelists began the training for the DCM method in the afternoon, but one panelist did not complete the session and thus did not contribute to the setting of the final passing score due to a scheduling conflict; therefore, for some items that panelist is listed as "no response."

Table 4.32

Evaluation Survey Results: Direct Consensus Method

Question number and topic	Rating category	Panel	
		A (ICM first) ^a	B (DCM first)
3. Impression of training	Appropriate	100.0% (8)	100.0% (8)
	Somewhat appropriate	--	--
	Not appropriate	--	--
4. Length of time provided for setting passing score	About right	87.5% (7)	87.5% (7)
	Too little time	--	12.5% (1)
	Too much time	--	--
	No response	12.5% (1)	--
5. Factors that influenced the passing score set (indicate all that apply)	Definition of minimally competent CPA	88.9% (8)	85.7% (6)
	Difficulty of test items	77.8% (7)	100.0% (7)
	Item statistics	33.3% (3)	57.1% (4)
	Other panelists	44.4% (4)	42.9% (3)
	My experience in the field	66.7% (6)	85.7% (6)
	Knowledge and skills measured by test items	66.7% (6)	71.4% (5)
	Other (see open-ended question section)	11.1% (1)	14.3% (1)

(continued on next page)

Table 4.32 (continued)

Question number and topic	Rating category	Panel	
		A (ICM first) ^a	B (DCM first)
7. Comfort with level of participation in group discussions	Very comfortable	75.0% (6)	75.0% (6)
	Somewhat comfortable	--	25.0% (2)
	Unsure	12.5% (1)	--
	Somewhat uncomfortable	--	--
	Very uncomfortable	--	--
	No response	12.5% (1)	--
8. Belief that passing score set will be correctly placed on the score scale	Definitely yes	25.0% (2)	--
	Probably yes	50.0% (4)	87.5% (7)
	Unsure	12.5% (1)	--
	Probably no	--	12.5% (1)
	Definitely no	--	--
	No response	12.5% (1)	--
9. Training	1—Not at all clear	--	--
	2	--	--
	3	--	--
	4	37.5% (3)	25.0% (2)
	5—Clear	62.5% (5)	75.0% (6)

(continued on next page)

Table 4.32 (continued)

Question number and topic	Rating category	Panel	
		A (ICM first) ^a	B (DCM first)
10. Discussion of item ratings following completion of first set of individual ratings	1—Not at all useful	--	--
	2	--	--
	3	--	--
	4	25.0% (2)	37.5% (3)
	5—Useful	75.0% (6)	62.5% (5)
11. Item statistical information	1—Not at all helpful	--	--
	2	--	--
	3	37.5% (3)	--
	4	12.5% (1)	12.5% (1)
	5—Helpful	50.0% (4)	87.5% (7)
12. Level of confidence in DCM ratings	1—Very low	--	--
	2	--	--
	3	12.5% (1)	--
	4	25.0% (2)	62.5% (5)
	5—Very high	50.0% (4)	37.5% (3)
	No response	12.5% (1)	--

(continued on next page)

Table 4.32 (continued)

Question number and topic	Rating category	Panel	
		A (ICM first) ^a	B (DCM first)
13. Level of confidence in final passing score	1—Very low	--	--
	2	--	--
	3	12.5% (1)	12.5% (1)
	4	25.0% (2)	50.0% (4)
	5—Very high	50.0% (4)	37.5% (3)
	No response	12.5% (1)	--

Note. ICM = Item Cluster method; DCM = Direct Consensus method.

^aEight panelists began the training for the DCM method in the afternoon, but one panelist did not complete the session and thus did not contribute to the setting of the due to a scheduling conflict; therefore, for some items that panelist is listed as “no response.”

Table 4.33

Evaluation Survey Results: Item Cluster Method

Question number and topic	Rating category	Panel	
		A (ICM first)	B (DCM first)
14. Impression of training	Appropriate	87.5% (7)	62.5% (5)
	Somewhat appropriate	12.5% (1)	37.5% (3)
	Not appropriate	--	--
15. Length of time provided for setting passing score	About right	50.0% (4)	25.0% (2)
	Too little time	37.5% (3)	62.5% (5)
	Too much time	12.5% (1)	12.5% (1)

(continued on next page)

Table 4.33 (continued)

Question number and topic	Rating category	Panel	
		A (ICM first)	B (DCM first)
16. Factors that influenced the passing score set (indicate all that apply)	Definition of minimally competent CPA	100.0% (9)	85.7% (6)
	Difficulty of test items	88.9% (8)	100.0% (7)
	Item statistics	77.8% (7)	--
	Other panelists	22.2% (2)	28.6% (2)
	My experience in the field	66.7% (6)	71.4% (5)
	Knowledge and skills measured by test items	55.6% (5)	85.7% (6)
	Pattern of right and wrong answers across test items	33.3% (3)	57.1% (4)
	Number of correct answers given by the candidate	66.7% (6)	85.7% (6)
	Number of answer choices to test items	11.1% (1)	14.3% (1)
	Other (see open-ended question section)	--	14.3% (1)

(continued on next page)

Table 4.33 (continued)

Question number and topic	Rating category	Panel	
		A (ICM first)	B (DCM first)
17. Use made of incorrect answer information in rating candidate profiles	Considerable use	62.5% (5)	62.5% (5)
	Some use	37.5% (3)	37.5% (3)
	Limited use	--	--
	No use	--	--
19. Comfort with level of participation in group discussions	Very comfortable	75.0% (6)	50.0% (4)
	Somewhat comfortable	12.5% (1)	50.0% (4)
	Unsure	--	--
	Somewhat uncomfortable	12.5% (1)	--
	Very uncomfortable	--	--
20. Belief that passing score set will be correctly placed on the score scale	Definitely yes	12.5% (1)	--
	Probably yes	37.5% (3)	37.5% (3)
	Unsure	50.0% (4)	50.0% (4)
	Probably no	--	12.5% (1)
	Definitely no	--	--

(continued on next page)

Table 4.33 (continued)

Question number and topic	Rating category	Panel	
		A (ICM first)	B (DCM first)
21. Training	1—Not at all clear	--	--
	2	--	--
	3	--	12.5% (1)
	4	62.5% (5)	37.5% (3)
	5—Clear	37.5% (3)	50.0% (4)
22. Practice exercise	1—Not at all useful	--	--
	2	--	12.5% (1)
	3	--	12.5% (1)
	4	50.0% (4)	12.5% (1)
	5—Useful	50.0% (4)	62.5% (5)
23. Discussion of item ratings following completion of first set of individual ratings	1—Not at all useful	--	--
	2	--	--
	3	--	25.0% (2)
	4	50.0% (4)	50.0% (4)
	5—Useful	50.0% (4)	25.0% (2)

(continued on next page)

Table 4.33 (continued)

Question number and topic	Rating category	Panel	
		A (ICM first)	B (DCM first)
24. Item statistical information	1—Not at all helpful	--	--
	2	--	25.0% (2)
	3	--	37.5% (3)
	4	25.0% (2)	12.5% (1)
	5—Helpful	75.0% (6)	25.0% (2)
25. Level of confidence in item cluster ratings	1—Very low	--	--
	2	--	--
	3	62.5% (5)	42.9% (3)
	4	25.0% (2)	57.1% (4)
	5—Very high	12.5% (1)	--

Note. ICM = Item Cluster method; DCM = Direct Consensus method.

Table 4.34

Evaluation Survey Results: Open-Ended Questions

Question number and topic	Panel	
	A (ICM first)	B (DCM first)
5. Other factors that influenced passing score set	“Whether particular questions tested either (1) accounting theory or (2) accounting issues known to be covered and emphasized in academic accounting environment” (panelist A8)	“My recollection of whether the item was covered in CPA review courses” (panelist B4)

Note. Panelist number is given immediately after each comment so that responses within this table can be compared for a given panelist (however, since responses to the survey were anonymous, these panelist numbers are arbitrary and do not correspond to those given in results showing cut scores calculated for each panelist). Some of the questions for which open-ended responses are given below are also summarized in previous tables. In those cases, the panelist’s ratings is given in parentheses after the panelist number.

(continued on next page)

Table 4.34 (continued)

Question number and topic	Panel	
	A (ICM first)	B (DCM first)
6. How could the DCM training have been improved?	<p>“I feel it was just fine” (panelist A1)</p> <p>“Timing was good—questions seemed to repeat with morning so it was easier to answer (might have skewed “easy” factor)” (panelist A2)</p> <p>“Greater emphasis on various areas, more technical basis, to be tested and passed to be considered a minimally competent CPA” (panelist A4)</p> <p>“I believe the global mod process, whereby panelists had the option of increasing or decreasing their <u>overall</u> score, is open to some criticism because it does cause some panelists to change their score based on the “desire” to conform with the total panel’s wishes. I believe it would be advisable to make this process more private and allow the panelist to make their change without announcing it to the group” (panelist A5)</p>	<p>“It seemed appropriate” (panelist B9)</p> <p>“Make more clear that discussion is not intended to determine easy vs. too hard questions, rather to persuade participants toward ‘group’ consensus” (panelist B2)</p> <p>“Explain in the beginning that purpose is to come to 1 passing grade, not an ‘averaged’ consensus” (panelist B3)</p> <p>“I think the training was informative and appropriate. Questions were asked by panel and relevant answers were given in response” (panelist B5)</p> <p>“I think a full day is necessary so that more questions can be reviewed and more detailed discussion can be discussed” (panelist B6)</p> <p>(continued)</p>

(continued on next page)

Table 4.34 (continued)

Question number and topic	Panel A (ICM first)	Panel B (DCM first)
6. (continued) How could the DCM training have been improved?		<p>“While you are asking for what #s we rated, have participants circle their ratings that are not in line with the group. Then give them quiet time (5 min) to come up with a defense for their ratings, b/c there are too many factors to consider” (panelist B7)</p>
8. Explain your answer to the question of whether the passing score set using the DCM will be correctly placed on the exam score scale.	<p>“Don’t really know the final outcome as we were one of first panels” (panelist A1; rating of “unsure”) “I liked this method the best—it sat better with me” (panelist A2; rating of “definitely yes”) “The range of questions appeared to be covering enough of a broad area to be beneficial” (panelist A3; rating of “probably yes”) (continued)</p>	<p>“I thought we reached a pretty solid passing score that would in fact be representative of the ‘minimally competent CPA’” (panelist B3; rating of “probably yes”) “My guess is that we were too strict given that we likely have higher expectations of CPAs than the minimally competent CPA” (panelist B4; rating of “probably no”) (continued)</p>

(continued on next page)

Table 4.34 (continued)

Question number and topic	Panel
	<p style="text-align: center;">A (ICM first)</p> <p style="text-align: center;">B (DCM first)</p>
<p>8. (continued)</p> <p>Explain your answer to the question of whether the passing score set using the DCM will be correctly placed on the exam score scale.</p>	<p>“Our group was able to come to a clear consensus” (panelist A4; rating of “probably yes”)</p> <p>“Subject to the undue influence of the overall group criticism that I raised on the previous page” (panelist A5; rating of “probably yes”)</p> <p>“The passing score set by this panel, I believe was a little low. I don’t believe in making it too easy for the candidates. Candidates should be required to possess a certain level of general knowledge. At some point in their career it is very likely the issues will be come across” (panelist A6; rating of “probably yes”)</p> <p>“I believe that the % is reasonable and supported by the members of Panel B. I do not say definitely because other regions of the country have not been compared and my panel was limited to a limited # of questions for review” (panelist B6; rating of “probably yes”)</p> <p>“For the most part, the group agreed on many of the questions. There were only a few deviants” (panelist B7; rating of “probably yes”)</p> <p>“Based upon our group study, I believe that the direct consensus method should be fair. There is potential for some on the panel to be influenced by others” (panelist B8; rating of “probably yes”)</p>

(continued on next page)

Table 4.34 (continued)

Question number and topic	Panel	
	A (ICM first)	B (DCM first)
16. Other factors that influenced passing score set	<p>“Whether I believed the wrong answer was ‘unacceptable’ (i.e. <u>way off base</u>) or had some merit” (panelist B4)</p>	

(continued on next page)

Table 4.34 (continued)

Question number and topic	Panel A (ICM first)	Panel B (DCM first)
18. How could the ICM training have been improved?	<p>“N/A” (panelist 1)</p> <p>“Too much time spent on problems—but I guess that is the basis of method” (panelist A2)</p> <p>“I did not feel comfortable with this method. I felt that the intent was to determine a ‘weighted average’ of correct items with greater consideration given to answering particular questions correctly. However, in the end, the overall # of correct items was the determining factor and my particular discussion or determination of particular question importance was virtually ignored. I believe the theory of this method would be useful, however, the practical application of this method would not be feasible. I was uncomfortable with this method determining a passing score” (panelist A4)</p> <p>(continued)</p>	<p>“Fewer individuals with more stark contrasts. It got tedious—people can zone out and not look for distinctions” (panelist B2)</p> <p>“If the individuals were scattered for each cluster (e.g., 2 right answer, 5 right answers, 1 right answer, etc.). The way it was set up you tended to get into a rhythm. By scattering the scores down the cluster you may have gotten different results from the panel” (panelist B3)</p> <p>“If there were less profiles to analyze I felt it would have been more productive” (panelist B5)</p> <p>“More time to complete each cluster and discuss results” (panelist B6)</p> <p>(continued)</p>

(continued on next page)

Table 4.34 (continued)

Question number and topic	Panel	
	A (ICM first)	B (DCM first)
18. (continued) How could the ICM training have been improved?	<p>“The definitions as to candidate ratings need to be improved. Either the number of ratings needs to be changed, or the definition of above/below borderline should be clarified more in the training.” (panelist A5)</p> <p>“More time to make decisions or ratings and having open forum to discuss” (panelist A7)</p>	<p>“It seemed appropriate” (panelist B1)</p> <p>“Have participants label the question #s that they thought were hard. Then have them circle all the participants who gave ridiculous answers (based on the question info)” (panelist B7)</p> <p>“If incorrect information is to be used, calculations should be shown to grader” (panelist B8)</p>

(continued on next page)

Table 4.34 (continued)

Question number and topic	Panel	
	A (ICM first)	B (DCM first)
20. Explain your answer to the question of whether the passing score set using the ICM will be correctly placed on the exam score scale.	<p>“We really didn’t get to a final score” (panelist A1; rating of “unsure”)</p> <p>“I wasn’t as comfortable with this method—too minute to detail questions. Would rather look at overall clusters” (panelist A2; rating of “probably yes”)</p> <p>“The knowledge base of the panelists appeared to be adequate” (panelist A3; rating of “probably yes”)</p> <p>“A particulate score was never set, therefore, I am unsure what passing score could be utilized” (panelist A4; rating of “unsure”)</p> <p>“Our panel was unsure as to what was the final passing score—several panelists expressed confusion” (panelist A5; rating of “unsure”)</p>	<p>“People were using different criteria to do ratings” (panelist B2; rating of “probably yes”)</p> <p>“I am not really sure how what we did actually established a passing score” (panelist B3; rating of “probably yes”)</p> <p>“Again, I bet that we were too strict for what is expected of the minimum competent CPA” (panelist B4; rating of “probably no”)</p> <p>“I’m not as comfortable with this method. It appears to be more judgmental” (panelist B6; rating of “unsure”)</p> <p>“It seems fair rating candidates against one another and based on how well they answered the questions (i.e., even if they go the answer wrong, they gave the next best answer, then they were above the borderline” (panelist B7; rating of “probably yes”)</p>

(continued on next page)

Table 4.34 (continued)

Question number and topic	Panel
26. Explain your answer to the question of which method you think will produce defensible passing scores for the CPA Exam.	<p style="text-align: center;">A (ICM first)</p> <p style="text-align: center;">B (DCM first)</p>
	<p>“It’s not that I don’t have confidence in the Item Cluster Method but I think that the Direct Consensus Method produced better results. Plus it was much easier and less time consuming to apply” (panelist A2; indicated DCM only)</p>
	<p>“Confidence in both—Consensus probably more meaningful—Item cluster, it appeared that participants got too understanding of mistakes (panelist B2; indicated both methods)</p>
	<p>“The direct consensus method was easier to apply, but we had more info for the item cluster method and given the appropriate time it may actually produce a better passing score, because to some extent it takes away the effect of lucky guesses” (panelist B3; indicated both methods)</p>
	<p>“Both methods appear to break down what the panelists believe is a minimally competent CPA” (panelist A3; indicated both methods) (continued)</p>
	<p>“I have more confidence in the Direct Consensus Method because the interaction of the group was more integral. Panelists had differing views and had to defend or propose them” (panelist B5; indicated both methods) (continued)</p>

(continued on next page)

Table 4.34 (continued)

Question number and topic	Panel	
	A (ICM first)	B (DCM first)
<p>26. (continued)</p> <p>Explain your answer to the question of which method you think will produce defensible passing scores for the CPA Exam.</p>	<p>“(See answer to item #18) I also feel that the Direct Consensus Method allowed the experts the opportunity to better determine which questions would have more ‘weight’, which would be considered in determining a passing score, which questions represented knowledge a minimally competent CPA would have and utilize same in determining a passing score” (panelist A4; indicated DCM only)</p> <p>“As denoted(?) in the training for the Item Cluster Method, I did not have confidence as to what was our determination of a passing score” (panelist A5; indicated DCM only)</p> <p>“The direct consensus method forced me to evaluate the difficulty of the question. I paid more attention to the issues the question was addressing. I evaluated where the candidate might get hung up, if the wording was intimidating” (panelist 6; indicated DCM only)</p>	<p>“It appeared to give a more number specific result” (panelist B6; indicated DCM only)</p> <p>“Both seemed fair methodologies” (panelist B7; indicated both methods)</p> <p>“The item cluster method seemed more subjective and it seemed that I relied on the % correct to guide my assessments (which I don’t think should have impacted my decisions) in this method” (panelist B1; indicated DCM only)</p>

(continued on next page)

Table 4.34 (continued)

Question number and topic	Panel	
	A (ICM first)	B (DCM first)
28. Do you think that your participation in the first session influenced your ratings in the second session?	<p>“Not really—but like I said some questions repeated. Therefore I don’t know if those questions became easier (i.e. set standard higher for minimally competent CPA)” (panelist A2)</p> <p>“Yes. I think knowing what the other panelists thought after the first session can influence the answers” (panelist A3)</p> <p>“Not at all, other than that some of the questions from the 1st session were repeated in the second session, which skewed the perceived difficulty of the questions” (panelist 5)</p>	<p>“No...not really” (panelist B4)</p> <p>“No, I took each session on its own merit” (panelist B6)</p>

(continued on next page)

Table 4.34 (continued)

Question number and topic	Panel
	<div data-bbox="260 1066 281 1093">A</div> <div data-bbox="294 1011 317 1152">(ICM first)</div>
	<div data-bbox="264 407 285 435">B</div> <div data-bbox="298 344 321 497">(DCM first)</div>
29. Explanation of which one method should be chosen by AICPA to set passing scores.	<p>“I would suggest the Direct Consensus Method for the reasons discussed above” [see question 26] (panelist A2)</p> <p>“Direct consensus method. This method appears to be more subjective to what a CPA would do” (panelist A3)</p> <p>“Direct consensus method—see answers to items #18 and 26” (panelist A4)</p> <p>“Direct consensus method—I believe emphasis on difficulty and ability level needed for the questions and problems, rather than the more subjective ‘borderline’ assessment of the minimally competent candidate, is the more defensible approach” (panelist A5)</p> <p>(continued)</p>
	<p>“Consensus—see # 26” (panelist B2)</p> <p>“Item cluster. It gives more credit to those who answer a question with a reasonable possibility rather than something that is way off base” (panelist B4)</p> <p>“Direct Consensus—see question 26” (panelist B5)</p> <p>“I would choose Direct Consensus Method. I believe it produces a more concrete(?) scoring” (panelist B6)</p> <p>“Item Cluster Method because it compares candidates against each other” (panelist B7)</p> <p>“Direct consensus because I feel that it is less subject to bias” (panelist B8)</p>

(continued on next page)

Table 4.34 (continued)

Question number and topic	Panel
	A (ICM first) B (DCM first)
<p>29. (continued)</p> <p>Explanation of which one method should be chosen by AICPA to set passing scores.</p>	<p>“Direct Consensus Method—reasons explained in #26. Item Cluster does not force you to factor in whether they guessed or really if the question is reasonable to a minimally competent CPA” (panelist A6)</p> <p>In the item cluster method, we analyzed individual questions in more detail, which seemed to be more accurate (because we focused more on the details, not on obtaining consensus). However, Consensus did bring more themes and objectives, which is also important” (panelist A8)</p>

(continued on next page)

Table 4.34 (continued)

Question number and topic	Panel A (ICM first)	Panel B (DCM first)
30. Additional comments	<p>“Very good panel discussion; enjoyed participating” (panelist A1)</p> <p>“It was useful to keep the definition of the minimally competent CPA in mind when applying the methods—however it is not a difficult subject to comprehend. Therefore I would limit the ‘re-learning’ of this concept throughout the day. Maybe a reminder to ‘remember the definition of a minimally competent CPA’ would be more appropriate. Other than that, it was an interesting day remembering this exam.” (panelist A2)</p> <p>(continued)</p>	<p>“Statistical info—%s on correct answers should be bold to facilitate identifying that score” (panelist B2)</p>

(continued on next page)

Table 4.34 (continued)

Question number and topic	Panel
	<div data-bbox="260 1083 326 1172">A (ICM first)</div> <div data-bbox="260 358 326 515">B (DCM first)</div>
30. (continued)	
Additional comments	
	<p>“I didn’t feel 100% comfortable in my participation for the following reasons: (1) I am not involved in FARE at work; (2) I am 3 years removed from studying and knowing FARE; (3) I would probably be a better panelist for ARE because of my experience in governmental accounting” (panelist A3)</p> <p>“I think the focus panelists composed of CPAs with 4–7 years experience, and supervisory experience is an excellent method. As stated during the panel discussion, the groups should be expanded to more panels and should cover all topics of the CPA Exam” (panelist A4)</p> <p>(continued)</p>

(continued on next page)

Table 4.34 (continued)

Question number and topic	Panel	
	A (ICM first)	B (DCM first)
30. (continued)		
Additional comments	<p>“After doing both methods I felt that I personally did not concentrate enough on whether the questions are appropriate for a minimally competent CPA. The direct consensus method zoned in on whether the candidates would be able to answer the questions. I considered the difficulty, relevance and wording of the questions” (panelist A6)</p>	<p>“The process was very interesting and I would be pleased participate in the future” (panelist A8)</p>

CHAPTER 5

DISCUSSION

Within this section, the results of the study will be discussed within the context of the three types of validity evidence outlined in Table 2.3. Conclusions and directions for future research are also presented.

5.1 Evaluation of Results Within Validity Framework

Kane (1994, 2001) presented a framework wherein results of standard-setting processes can be evaluated according to three general criteria—procedural, internal, and external. Within each of these three areas, additional subcriteria can be identified. Each of these three sources of validity evidence will be discussed in terms of the two methods used in this study. This approach will allow integration of the results presented in Chapter 4.

5.1.1 Procedural Evidence of Validity

According to Kane (2001), “the fact that a standard setting study has employed an apparently sound procedure in a thorough and systematic way, and has where possible, included various checks on the consistency and reasonableness of the results encourages us to have faith in the results” (p. 68). Kane noted that the reasonableness of procedures is often the primary source of evidence, and that policy decisions based on resulting cut scores are viewed with more confidence if procedures are followed in a sound manner by panelists who are qualified and understand the process. Although the current study afforded the opportunity to replicate results both within and across

methods, which is a luxury not often present in operational standard-setting studies, procedural evidence is still a critical component in filling out the validity picture. Procedural evidence will be grouped into the five areas contained in Table 2.3: explicitness, practicability, implementation of procedures, panelist feedback, and documentation. Panelist feedback will be discussed within other areas as it facilitates their evaluation (i.e., how practicable the panelists felt each method was), as well in a separate section.

Explicitness. Van der Linden (1995) defined explicitness as the degree to which the standard-setting process was clearly and explicitly defined before implementation. One justification given by van der Linden for this criterion is that were it not met, the results of the standard-setting research would not be able to be communicated in a clear and meaningful manner. Additionally, however, he noted that it would be very difficult to apply the other validity criteria if the groundwork for the standard-setting process were not applied in a thorough fashion. In the current study, both methods were outlined clearly and in detail due to their being part of a well-scrutinized academic research effort. Though the explicitness criterion is a valuable one, it is perhaps more relevant to standard-setting efforts that are applied in an operational setting with less independent oversight of the process.

Practicability. Berk (1986) noted that technical defensibility of a standard-setting procedure is not sufficient; the method must also be capable of being implemented without great difficulty, data analysis must be feasible without laborious computations, and the procedures must be credible and interpretable to laypeople.

Because much of the evidence for this criterion can be gleaned from the panelists themselves, panelist feedback will be discussed within this section as well.

On the practicability criterion, the Direct Consensus method appears to have an edge. As reflected in the time needed for implementation, the Direct Consensus method is more streamlined than the Item Cluster method (see Table 4.28). Its relative simplicity results in less time being required for training, and data analyses required after completion of the study are minimal. In fact, no analyses are needed to arrive at the cut score; the only computations required are those related to characteristics of the panelist ratings (i.e., consistency, correspondence to empirical data).

Panelists appeared to find the Direct Consensus method more readily understandable than the Item Cluster method, and it is probably not a stretch to postulate that laypeople would as well. The act of reviewing test items and coming to a consensus on the number that an examinee would answer correctly is a much simpler task than reviewing numerous examinee profiles and assigning a value on a six-point scale to them. Panelists also appeared to have some discomfort with the fact that no cut score was arrived at by the conclusion of the session with the Item Cluster method, which is linked to the more laborious nature of the computations necessary to estimate a cut score with that method. However, it should be noted that if time and resources so allowed, cut scores could be calculated at the session itself, perhaps most easily with the boundary method. Were this feature added to the Item Cluster method, discussion about the passing score could also be added to the procedure.

Returning to the practicability of the Item Cluster method, though the calculations required to estimate cut scores with the boundary, regression, and equating

approaches are not statistically complex, they are time-consuming. When combined with the amount of data manipulation required to document consistency of panelist ratings and their relationship to examinee performance, these computations might give pause to staff at a busy licensing or certification agency.

Implementation of procedures. According to Kane (1994), the extent to which the selection and training of panelists, definition of the performance standard, and data collection were implemented in a systematic and thorough fashion is an important source of procedural validity evidence. In the current study, panelists were for the most part appropriate for the task at hand, since they had supervised entry-level CPAs in the relevant accounting area in the recent past. However, two panelists (one on each panel) did not meet this criterion. In addition, the average years of experience of panel members was hired than originally desired. The number of panelists who served on each panel (eight panelists for three of the sessions; seven for the other) was adequate. Cizek (1996b) observed that Livingston and Zieky (1982) described studies conducted with as few as five panelists, but that Smith, Smith, Richards, and Barnhardt (1988, as cited in Cizek, 1996b) found that ratings were still quite variable even with 10 panelists. Ideally, more panelists would have participated, but as is often the case, it was difficult to find professionals who could afford the time away from the office, despite the provision of continuing professional education credits. The fact that one panelist in Panel A left shortly into the afternoon session was unfortunate, since the panels were then unbalanced in number in that session.

Training of the panelists is also critical to the sound implementation of a standard-setting method (Cizek, 1996b; Kane, 1994). Training was less complicated

and time-consuming for the Direct Consensus method than for the Item Cluster method, since the latter method is far more complex operationally. Even though more time was devoted to training for the Item Cluster method (see Tables 4.29 and 4.30), the time frame within which the study was conducted—one day with two sessions—most likely caused training to be more abbreviated than it should have been. Panelists' responses to the training-related questions in the Evaluation Survey may reflect this, since they appeared to view training for the Direct Consensus method (Table 4.32, questions 3 and 9) slightly more favorably than that for the Item Cluster method (Table 4.33, questions 14 and 21). On this facet of the procedural criterion, the Direct Consensus method appears to have a slight edge.

A definition of the performance standard—i.e., the description of the minimally competent CPA—was provided to panelists within a 15-minute period prior to the separation of the group into two panels (see Table 3.3). A description of the minimally competent CPA that had been previously adopted by the AICPA Board of Examiners was distributed for review and discussion. Panelists thus did not participate in the definition, but were led through a discussion of its features. Panelists were also given a copy of the description in their method-specific packet of materials, and were asked to keep this description available for easy reference throughout the day. Panelists' responses to question 1 in the evaluation survey (see Table 4.31) indicate that nine of the panelists (about half of them) felt that the training was "very clear." Six panelists felt that it was "clear," and one, "somewhat clear." In the additional comments question in the open-ended question section of the survey (Table 4.34, question 30), two panelists provided divergent feedback regarding the definition of the minimally

competent CPA. Panelist A2 thought that there was too much “re-learning” of the concept during the sessions, while Panelist A6 thought that he or she did not concentrate enough on the notion of the minimally competent CPA (as indicated in the note to this table, these panelist numbers do not correspond to those in the tables of results, since evaluation survey responses were confidential). Overall, the definition of the performance standard appears to have been adequate, though more time should have perhaps been allotted had the schedule allowed it.

Kane (1994) noted that the procedures used to collect the standard-setting data should be systematic and accurate, and provided suggestions for improving the data’s quality. Those ideas included having panelists provide ratings more than once, which was accomplished in both methods considered in the current study. The provision of empirical performance data is viewed by Kane and others (e.g., Jaeger, 1982, 1989) as being helpful as well. Again, this was a feature of both methods in the study. In addition, discussion among panelists is seen as facilitating the setting of cut scores at reasonable levels. While both methods had a discussion component, it was more critical to the Direct Consensus method because of its consensus-building nature. For the afternoon session implementation of the Item Cluster method, discussion was more limited than in the morning session, either due to a facilitator effect, fatigue, or the personalities of the panelists.

Panelist feedback. Some panelist feedback has been included in previous subsections as applicable. However, it is also informative to consider panelists’ responses to those questions on the evaluation survey that ask directly about the methods, as well as panelists’ answers to the open-ended questions. To question 26 (see

Table 4.31), which asked panelists to indicate which method would produce a defensible passing score for the exam, results were different for the two panels. For Panel A, which used the Item Cluster method first, five panelists indicated that only the Direct Consensus method would produce a defensible passing score, while two panelists indicated both methods (the remaining panelist did not complete the afternoon session and therefore did not answer this question). For Panel B, which used the Direct Consensus method first, two panelists indicated the Direct Consensus method only, while six panelists indicated both methods. Thus Panel A members were more critical of the Item Cluster method than Panel B members.

In terms of making a recommendation for which method should be used if only one could be chosen (Table 4.31, question 29), the results were more similar across panels. For Panel A, which used the Item Cluster method first, six panelists indicated the Direct Consensus method, and one panelist indicated both methods. For Panel B, which used the Direct Consensus method first, five panelists indicated the Direct Consensus method, two panelists the Item Cluster method, and one panelist both methods. Thus the majority of panelists would recommend the Direct Consensus method if only one method could be endorsed.

Panelists were also asked whether each method would result in a passing score that would be correctly placed on the score scale. For the Direct Consensus method (Table 4.32, question 8), results differed depending on whether the panel had implemented the method first or second, paralleling the trend for question 29 as reviewed above. In Panel A, which implemented the method second, two panelists indicated that the cut score would “definitely” be placed correctly, four panelists

indicated “probably,” and one was “unsure.” In Panel B, which implemented the method first, seven panelists indicated “probably,” and one “probably not.” For the Item Cluster method (Table 4.33, question 20), results were fairly similar across panels. In Panel A, which implemented the method first, one panelist indicated “definitely,” three panelists “probably,” and four panelists “unsure.” In Panel B, which implemented the method second, three panelists indicated “probably,” four panelists “unsure,” and one panelist “probably no.” Panelists appeared to believe that the passing score would be more likely to be set correctly on the scoring scale with the Direct Consensus method.

One other set of questions addressed the level of confidence that panelists felt about the cut score set (Direct Consensus method; Table 4.32, question 13) or about the ratings provided (Item Cluster method; Table 4.33, question 25). In Panel A, which implemented the Direct Consensus method second, four panelists indicated a “very high” level of confidence in the cut score set with that method. In Panel B, which implemented the method first, three panelists indicated that level of confidence. For the Item Cluster method, panelist indicated less confidence in the method, in this case as reflected in the item ratings and not the cut score. In Panel A, which implemented the method first, only one panelist indicated a “very high” level of confidence; in Panel B, no panelists did.

In general, panelists appeared to have more confidence in the Direct Consensus method and would recommend it if only one method could be chosen. While panelists’ opinions do not in and of themselves indicate the superiority of one method over another, they do serve as a valuable source of validity evidence.

Documentation. As with the explicitness criterion, in the current study documentation is assured because of the academic nature of the research. As a result, execution of both methods in this study fulfills this criterion.

5.1.2 Internal Evidence of Validity

Evidence to be discussed within this section includes consistency within method, and intrapanelist and interpanelist consistency, three of the areas outlined in Table 2.3. (The fourth area, other measures—the consistency of cut scores across item types, content areas, and cognitive processes—was not investigated in this study.)

Consistency within method. Perhaps the most important source of internal validity evidence provided in this study are the replications within method. Kane (1994, 2001) noted that the best way to estimate the standard error of the cut score is to convene different groups of panelists on the same or different occasions. In the current study, both methods were used by two different panels, on different test forms. The degree to which the cut scores are similar across these two implementations provides valuable information about the replicability of the cut score with a given method.

Cut scores from the Direct Consensus and Item Cluster methods are compared directly in Table 4.27. Cut scores obtained by the two panels with the Direct Consensus method were much farther apart than those resulting from the two panels' implementation of the Item Cluster method. The Direct Consensus method cut scores were 30 and 34. This difference of 4 points is much larger than those observed with the Item Cluster method, which depending on approach selected range from 0.68 to 1.79 points. However, since the cut scores are tied to test forms with slightly different

lengths (by one item), perhaps a more meaningful comparison is the percentage of items that would need to be answered correctly by an examinee in order to meet the cut score. For the Direct Consensus method, there is a 7% difference in these percentages. In contrast, the differences for the Item Cluster method, depending on approach chosen for comparison, range from 1% to 3%.

In general, then, the Item Cluster method appears to produce more consistent results across replications than the Direct Consensus method. However, this conclusion must be drawn only with a caveat regarding the different dynamics that arose during the two sessions in which the Direct Consensus method was implemented. In the morning session, the facilitators became aware of a belief on the part of some panel members that a reasonable expectation would be that examinees get approximately 75% of the items correct. Not surprisingly then, the cut score set by that panel resulting in 74% of the items needing to be answered correctly. Apparently, some panel members thought that this was an operational policy relevant to this exam. However, this belief didn't become apparent until later in the session, at which time it had already affected the cut scores set. In the afternoon session, facilitators' awareness of this belief arose and was addressed much earlier, probably impacting the resulting cut score to a lesser degree. In subsequent implementations of the Direct Consensus method, care should be taken to ensure that the transparent nature of the cut score and its relation to number of items correct does not lead to panelists' preconceptions unduly influencing the cut score.

An ancillary issue related to the Item Cluster method is perhaps best discussed in this section, since it in a sense relates to consistency within method. That issue is which calculation approach to use to arrive at the cut score. Cut scores obtained using

the different approaches are shown in Tables 4.7 and 4.8. When the panel means of panelist cut scores are compared within a session, the range in cut scores (i.e., the difference between the largest and smallest cut scores across calculation approaches) for both panels are very similar—0.70 points for Panel A (morning), and 0.74 for Panel B (afternoon). The boundary method cut score is lowest for both panels. The highest cut score was obtained by the equating method for Panel A, and the linear regression method for Panel B. However, the differences are so small that such contrasts are likely without merit.

The small size of the differences suggest that the approach for calculating the Item Cluster method cut score be chosen on practical and theoretical grounds. The boundary and equating approaches are both fairly easy to implement, both at the panelist and panel level. However, the regression approach is more labor-intensive, particularly at the panelist level. This suggests that the regression approach be chosen only if it is clearly superior on a theoretical basis. At this point the number of panelists, and thus the size of the data set, may come into play. If the number of data points (i.e., ratings) is small, perhaps a model-based approach such as regression would be the best choice, since the very small number of ratings that are on the boundary would then not unduly influence the cut score, resulting in a more stable estimate. However, if the number of data points is large, it could be argued that the equating or boundary approaches are to be preferred, since they focus in on the area of the scale most relevant to the task at hand. Given a fairly large number of panelists, the equating approach may be judged more appropriate since regression artifacts will not be a factor. Given the high correlations between panelist ratings and examinee profile scores, those artifacts

are probably not an important factor in this study. Nonetheless, the equating method appears to be the most attractive choice.

Intrapanelist consistency. Evidence for this criterion is provided by the degree of relationship between each individual panelist's ratings and empirical data provided to them. Evaluation of this information differs based on the standard-setting method involved, so each will be discussed in turn.

In the Direct Consensus method, panelists were provided with *p*-values, or the percentage of examinees at the operational administration that got the item correct. This information was given to examinees only after the first round of ratings. Of interest is the degree to which there was a correspondence between ratings and *p*-values for both rounds. Tables 4.5 and 4.6 present the correlations between the mean *p*-value for the item cluster and panelist ratings, the latter represented by the percentage of items that panelists judged a borderline candidate as needing to answer correctly. Though this information is of interest, it is important to note that these correlations should be interpreted with caution, given that each is based on only five sets of data points, all of which are restricted in range. The four correlations—two for each panel, one for round one and one for round two—are all moderate. However, their meaningfulness must be questioned given that for Panel A (afternoon session) one change—a 0.04 increase in the mean panelist rating for cluster 5 from round one to round two—caused the correlation to drop from 0.64 to 0.30. In general, this information is of limited utility because of its nature.

For the Item Cluster method, there is much more data to use for analyses of the relationship between panelist ratings and actual examinee performance. In this method,

the correlations were calculated for each panelist by cluster, using two sets of data for each examinee profile: (1) the panelist's rating on the six-point scale and (2) the actual cluster score received by that examinee. In contrast to the limited data on which the Direct Consensus method correlations were based (one set of values for each cluster), for the Item Cluster method there were between 17 and 20 sets of values for each panelist for each cluster. The average correlation for each panel, for both rounds, was 0.95. For Panel A, individual panelist correlations ranged from 0.90 to 0.98 for Round One, and 0.88 to 0.98 for Round Two. For Panel B, individual panelist correlations ranged from 0.88 to 0.99 for both rounds. The correlations appear to reflect a reasonable degree of relationship between panelist ratings and actual examinee performance. Were those values closer to 1.00, it would be difficult to argue that panelists were actually performing the assigned task—reviewing the pattern of answers in an examinee profile—versus just basing their ratings on the total score for the profile. However, panelists' responses to question 16 on the evaluation survey (see Table 4.33) do cast some doubt on the degree to which they used the distractor information. Only three of the panelists on Panel A and four on Panel B indicated that the patterns of right and wrong answers were factors that influenced the passing score set.

Additional information related to intrapanelist consistency is provided by the degree to which panelists modified their ratings from round one to round two (and, for the Direct Consensus method, from round two to the global modification ratings). As noted earlier, these changes could reflect both the empirical data given after round one and any group processes that take place after review of round one ratings. Information

related to this aspect of intrapanelist consistency are presented in Tables 4.1 and 4.2 (Direct Consensus method) and 4.23 and 4.24 (Item Cluster method).

For the Direct Consensus method, more changes were made between round two and the global modification ratings than between round one and round two. It is interesting to note that three panelists (B6, B8, and A1) made no changes at all; this may suggest that these panelists either did not review empirical information or allow themselves to be influenced by consensus-building discussions. For the Item Cluster method, the number of changes made between rounds varied greatly between panels. In Panel A (morning session), only one panelist made no changes, and the mean number of changes was 10, with the total number of possible changes being 96 (the total number of examinee profiles). In Panel B, however, five panelists made no changes, and the mean number of changes was one. It is not clear whether this disparity is due to the difference in facilitators (the facilitator for the morning session was more experienced both overall and with this method) or due to panel-specific factors. The latter could be due either to difference in that panel's nature or to a fatigue factor caused by implementing the more complicated standard-setting procedure later in the day. But it is of interest that although the number of changes varied greatly from Panel A to B, the resulting cut scores were still very similar.

Interpanelist consistency. Evidence for this criterion is provided by the degree to which ratings were consistent across panelists. Perhaps the most helpful information for this criterion is presented in Tables 4.1 and 4.2 (Direct Consensus method) and Tables 4.13 to 4.22 (Item Cluster method). Information presented in the Direct Consensus method tables suggests that panelists were fairly consistent in their ratings.

For Panel A, the round one cut scores ranged from 31 to 35; for round two, from 31 to 36; and for the global modification cut scores, from 33 to 35. The only panelist that appeared to be slightly out of line with the rest was panelist B6, who set the highest cut score (35) and did not change it from round one; interestingly, this was one of the panelists who had not supervised entry-level CPAs in the past two years. Similarly, panelist B8 set the next highest cut score (34) and did not change it. However, since the resulting consensus cut score was 34, neither of these panelists are aberrant in the sense of being far afield from what the panelists came to agree on as reasonable. For Panel B, the round one cut scores ranged from 27 to 31; for round two, from 28 to 31; and for the global modification cut scores, from 29 to 30. None of the panelists appeared to be far apart from the panel as a whole with their ratings. The one panelist who did not revise his or her cut score at all was panelist A1, whose cut score matched the resulting consensus cut score of 30.

For the Item Cluster method, inspection of Tables 4.13 to 4.22 reveals that for the majority of examinee profile scores, panelists did not differ more than one rating point from each other. At the cut score level, the difference between minimum and maximum panelist cut scores ranged, depending on calculation method, from 1.97 to 3.31 for Panel A (morning; Table 4.7), and from 2.33 to 4.43 for Panel B (afternoon; Table 4.8). For both panels, the cubic regression approach resulted in the greatest spread of panelist cut scores.

5.1.3 External Evidence of Validity

Evidence to be discussed within this section includes comparisons between standard-setting methods and evaluation of reasonableness of the cut scores, two of the

areas outlined in Table 2.3. (The third area, comparisons to other sources of information, was not investigated in this study.)

Comparisons between methods. The current study offers the opportunity to directly compare the results of two standard-setting methods, both of which were conducted with the same panelists and the same test form. As shown in Table 4.27, the Direct Consensus method, while not yielding consistent cut scores within the method, produced cut scores that were in all cases higher than those produced by the Item Cluster method. Perhaps the most informative comparison between the two methods is that of the cut scores obtained when each method was the first one to be applied by the panel. Interestingly, this session was also the one in which the cut scores differed most between the methods—from 5.85 to 6.55 points depending on the Item Cluster method of calculation chosen for comparison. Unfortunately, this is also the session in which the prior beliefs of the panelists may have most strongly impacted the cut score, since they appeared to have thought they should set a cut score that reflected examinees' getting 75% of the items correct. In the second session, where that issue did not appear to impact panelists' ratings to as great a degree, the difference between methods was smaller, ranging from 3.23 to 3.71 points depending on the Item Cluster method chosen for comparison. It is of interest to note that in a previous study in which the Item Cluster method was implemented (Mills et al., 2000), it also yielded cut scores lower than the other method—in that case, the Angoff method.

Reasonableness of cut scores. Evidence relating to the reasonableness of the standards obtained with both methods is provided in this study by impact data, or the percentage of examinees estimated to pass the exam using the cut score. However,

evaluation of these impact data should be tempered by the fact that these estimates were obtained via an estimation process in which several assumptions were made, such as the shape of the score distribution. Nonetheless, inspection of these estimated pass rates reinforce the impression provided by the cut scores themselves—that the two standard-setting methods produced quite different results.

As shown in Table 4.27, estimated pass rates for the Direct Consensus method were 27% for Panel B (morning session) and 35% for Panel A (afternoon session), for an average method pass rate of 31%. Estimated pass rates for the Item Cluster method ranged from 49% to 52% for the morning session depending on approach used to analyze the data, and from 48% to 50% for the afternoon session, for an average method pass rate of 50%.

Across methods, the differences are quite striking. Using a conservative estimate of the pool for this exam of approximately 40,000 examinees (see, e.g., Pitoniak, Sireci, & Luecht, in press), the 19% difference in average method cut scores across methods (50% minus 31%) would result in 7,600 more examinees passing with the Item Cluster cut score than with the Direct Consensus cut score. Even within a method, apparently small differences across sessions can affect a surprisingly large number of examinees. For example, the 2% difference for the Item Cluster method cut scores within the afternoon session would differentially affect 800 examinees.

As far as the reasonableness of any of the cut scores is concerned, a comparison may be made to the average pass rate for this section of the exam obtained operationally. Pass rates over recent administrations ranged from approximately 24% to 28% (B. Biskin, personal communication, November 2, 2001). These are clearly more

in line with the estimates obtained from the Direct Consensus method cut scores than those obtained with the Item Cluster method. However, comparisons to operational pass rates that have a link to a complicated history of policy decisions should be made with caution.

5.2 Conclusions

The current study provided valuable information about two new standard-setting methods. The Direct Consensus and Item Cluster methods yielded cut scores that were noticeably different from each other. It is of course impossible to know which standard-setting method came closest to the “true cut score,” since such a value does not exist. Standard-setting is a judgmental process, and as Kane (1994) noted, “there is no gold standard. There is not even a silver standard” (pp. 448–449). For that reason, we must rely upon the accumulated weight of the various sources of validity evidence outlined in the previous sections to judge the utility of each method.

Procedurally, the Direct Consensus method was preferred by panelists, due perhaps in part to its being less time consuming and easier to implement. The Item Cluster method is a more challenging approach that appears to demand more from panelists, and perhaps for that reason is less preferred by them. In addition, the method requires more time for the calculation of cut scores, a factor that may be important in licensing and certification applications.

Internal validity evidence suggests that for both methods panelists were consistent both within their own ratings and across the panels. The nature of the correlations between panelist ratings and examinee performance differed across

methods. For the Direct Consensus method, the correlations are of limited meaningfulness due to the small number of data points on which they are based, as well as the restriction of range in those points. For the Item Cluster method, the correlations are of more utility, and suggest that panelists' ratings bear a reasonable relationship to examinee performance, as reflected in profile scores.

In terms of external validity evidence, the Item Cluster method appears to produce more consistent results, an important consideration for those setting standards for licensure examinations. However, the cut scores yielded by the Item Cluster method were much lower than those obtained with the Direct Consensus method. Although as a result the estimated pass rates for Item Cluster cut scores are out of line with operational trends, caution must be exercised in making any firm conclusions on the appropriateness of those pass rates given policy issues that are naturally associated with setting operational cut scores.

Neither the Direct Consensus method nor the Item Cluster method can be ruled out on the basis of the validity evidence, a role that this evidence often plays most effectively (Kane, 1994) since a passing score can never technically be ruled in, or established as unassailable. However, the choice of which method to use in an operational setting should be guided by a consideration of each method's strengths and weaknesses. In a licensure and certification setting where time is a great concern, the Direct Consensus method might be preferred; however, care should be taken that the transparent relation between the cut score and number of items needing to be answered correctly does not unduly influence the resulting cut score. In a venue where more time is available for standard setting, the Item Cluster method may be a viable option.

However, the possibility that lower cut scores may be set with this method should be considered carefully before the method is implemented.

5.3 Future Research

Each of the methods evaluated in this study shows promise. The Direct Consensus method is preferred by panelists, and its ease of implementation, reflected in its requiring less time to conduct, make it an attractive option for licensing and certification applications, particularly in technology fields. However, the current study showed that care must be taken that the transparent connection between panelist ratings and percent-of items correct does not cause problems. Future studies should investigate the degree to which consistent cut scores can be set using this method.

The Item Cluster method has been shown in both the current study and in Mills et al. (2000) to set consistent cut scores that are consistent, but lower than other methods. In this study, the resulting estimated pass rates raise questions about the reasonableness of the cut scores set for this particular exam. One aspect of the method that should be investigated in future studies is the extent to which panelists use the examinee profile information provided to them. If the majority of panelists indicate, as they did in this study, that they did not use this information, the operational (as opposed to theoretical) fruitfulness of the method should be questioned.

APPENDIX A
SAMPLE ITEM RATING FORMS

Direct Consensus Method Rating Form

Name: _____

Note: Your task is to indicate the number of correct answers that the minimally competent candidate will produce. This number will be between 0 and the total number of items in the cluster.

Item Cluster	Number of Items in Cluster	Item Numbers	Estimated # Correct	
1	9	1, 3, 5, 7, 9, 11, 13, 17, 18	_____	_____
2	10	19, 21, 22, 23, 25, 27, 28, 29, 31, 33	_____	_____
3	7	34, 35, 37, 39, 41, 47, 49	_____	_____
4	9	50, 51, 53, 55, 57, 59, 61, 63, 65	_____	_____
5	10	78-87 (OOAF #3)	_____	_____
Total			_____	_____

**Item Cluster Method
Rating Form**

Practice Exercise

Name: _____

ID	Item 01	Item 02	Item 03	Item 04	Item 05	Score	Rating 1	Rating 2
001	B	D	A	1	A	1		
002	A	1	1	A	C	2		
003	B	1	C	A	1	2		
004	1	C	1	1	C	2		
005	A	1	B	1	1	3		
006	B	1	1	D	1	3		
007	B	1	1	1	1	4		
008	1	1	C	1	1	4		
009	1	1	A	1	1	4		
010	1	1	1	1	1	5		

Notes:

- Correct answers are indicated by "1."
- Incorrect answers are indicated by the answer selected by the candidate ("A" through "D").
- "Score" is equal to the number of "1s" in the row, which corresponds to the number of correct answers obtained by the candidate for this cluster.

APPENDIX B
EVALUATION SURVEY

Evaluation Survey for Uniform CPA Examination Standard-Setting Study

June 13, 2001

Your anonymous answers to the questions below will be used to evaluate the total standard-setting process and to suggest revisions in the process, where necessary. Thanks for your help in completing the evaluation. Please note the first section contains general questions related to the process. Then, the following sections are method-specific. You will answer questions about the Direct Consensus Method first, and the Item Cluster Method second. At the end of the form is a section with several additional questions. We would appreciate your filling out all sections.

General Questions

1. How clear were you with the description of the minimally competent CPA?
(Circle one)
 - a. Very Clear
 - b. Clear
 - c. Somewhat Clear
 - d. Not Clear

2. What Panel you were in? Panel name is listed on the label on the front of your folder. (Circle one)
 - a. Panel A
 - b. Panel B

Direct Consensus Method

3. What is your impression of the Direct Consensus Method training you received for setting a passing score? (Circle one)
- a. Appropriate
 - b. Somewhat Appropriate
 - c. Not Appropriate
4. How would you judge the length of time for setting a passing score with the Direct Consensus Method? (Circle one)
- a. About Right
 - b. Too Little Time
 - c. Too Much Time
5. What factors influenced the passing score you set with the Direct Consensus Method? (Circle all choices that apply.)
- a. The definition of a minimally competent CPA
 - b. The difficulty of the test items
 - c. The item statistics
 - d. Other panelists
 - e. My experience in the field
 - f. Knowledge and skills measured by the test items
 - g. Other (please specify: _____)
- _____
6. How could the Direct Consensus Method standard-setting training have been improved? (Continue on back if necessary)
- _____
- _____
- _____
- _____
- _____

7. How comfortable were you with your level of participation in group discussions? (Circle one)
- a. Very Comfortable
 - b. Somewhat Comfortable
 - c. Unsure
 - d. Somewhat Uncomfortable
 - e. Very Uncomfortable
8. Do you believe that the passing score set by the panel using the Direct Consensus Method will be correctly placed on the exam score scale? (Circle one)
- a. Definitely Yes
 - b. Probably Yes
 - c. Unsure
 - d. Probably No
 - e. Definitely No

Please explain your answer (continue on back if necessary):

For each of the statements below, please circle the rating that best represents your judgment about the Direct Consensus Method.

- | | | | | | |
|--|-------------------|---|---|---|---------|
| 9. The training for Direct Consensus Method was | 1 | 2 | 3 | 4 | 5 |
| | Not at all clear | | | | Clear |
| 10. The discussion of item ratings following completion of the first set of individual ratings was | 1 | 2 | 3 | 4 | 5 |
| | Not at all useful | | | | Useful |
| 11. The item statistical information was | 1 | 2 | 3 | 4 | 5 |
| | Not at helpful | | | | Helpful |

- | | | | | | |
|---|---------------|---|---|---|----------------|
| 12. My level of confidence in my Direct Consensus Method ratings is | 1
Very Low | 2 | 3 | 4 | 5
Very high |
| 13. My level of confidence in the final passing score is | 1
Very Low | 2 | 3 | 4 | 5
Very high |
-

Item Cluster Method

14. What is your impression of the Item Cluster Method training you received for setting a passing score? (Circle one)
- a. Appropriate
 - b. Somewhat Appropriate
 - c. Not Appropriate
15. How would you judge the length of time of this meeting for setting a passing score using the Item Cluster Method? (Circle one)
- a. About Right
 - b. Too Little Time
 - c. Too Much Time
16. What factors influenced the passing score you set with the Item Cluster Method? (Circle all choices which apply.)
- a. The definition of a minimally competent CPA
 - b. The difficulty of the test items
 - c. The item statistics
 - d. Other panelists
 - e. My experience in the field
 - f. Knowledge and skills measured by the test items
 - g. Pattern of right and wrong answers across test items
 - h. The number of correct answers given by the candidate
 - i. The number of answer choices to the test items
 - j. Other (please specify: _____)

17. How much use did you make of incorrect answer information in rating candidate profiles?
- a. Considerable Use
 - b. Some Use
 - c. Limited Use
 - d. No Use

18. How could the Item Cluster Method standard-setting training have been improved? (Continue on back if necessary)

19. How comfortable were you with your level of participation in group discussions? (Circle one)
- a. Very Comfortable
 - b. Somewhat Comfortable
 - c. Unsure
 - d. Somewhat Uncomfortable
 - e. Very Uncomfortable

20. Do you believe that the passing score set by the panel using the Item Cluster Method is correctly placed on the exam score scale? (Circle one)
- a. Definitely Yes
 - b. Probably Yes
 - c. Unsure
 - d. Probably No
 - e. Definitely No

Please explain your answer (continue on back if necessary):

For each of the statements below, please circle the rating that best represents your judgment about the Item Cluster Method.

- | | | | | | |
|--|---------------------------|---|---|---|-------------------|
| 21. The training for the item cluster method was | 1
Not at
all clear | 2 | 3 | 4 | 5
Clear |
| 22. The practice exercise with the item cluster method was | 1
Not at
all useful | 2 | 3 | 4 | 5
Useful |
| 23. The discussion of item ratings following completion of the first set of individual ratings was | 1
Not at
all useful | 2 | 3 | 4 | 5
Useful |
| 24. The item statistical information was | 1
Not at
helpful | 2 | 3 | 4 | 5
Helpful |
| 25. My level of confidence in the item cluster ratings is | 1
Very
Low | 2 | 3 | 4 | 5
Very
high |

Additional Questions

26. Which method do you think will produce defensible passing scores for the CPA Exam? (Circle one)
- a. I have confidence in both methods.
 - b. I only have confidence in the Direct Consensus Method.
 - c. I only have confidence in the Item Cluster Method.
 - d. I do not have confidence in either method.

Please explain your answer (continue on back if necessary):

27. Did the use of **two** methods for setting passing scores in the study cause you to become confused? (Circle one)
- a. Not at all
 - b. Occasionally, I was confused in applying the second method.
 - c. It definitely was a problem for me.

28. Do you think that your participation in the first session influenced your ratings in the second session? If so, please explain.

REFERENCES

- American Educational Research Association. (2000). AERA position statement concerning high-stakes testing in preK–12 education. Washington, DC: Author.
- American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1999). Standards for educational and psychological testing. Washington, DC: American Educational Research Association.
- American Institute of Certified Public Accountants. (1999). Setting performance standards on the May 1998 Uniform CPA Examination. New York: Author.
- Angoff, W. H. (1971). Scales, norms, and equivalent scores. In R. L. Thorndike (Ed.), Educational measurement (2nd ed., pp. 508–597). Washington, DC: American Council on Education.
- Angoff, W. H. (1984). Scales, norms, and equivalent scores. Princeton, NJ: Educational Testing Service.
- Angoff, W. H. (1988). Proposals for theoretical and applied development in measurement. Applied Measurement in Education, 1, 215–222.
- Bejar, I. I. (1983). Subject matter experts' assessment of item statistics. Applied Psychological Measurement, 7, 303–310.
- Berk, R. A. (1986). A consumer's guide to setting performance standards on criterion-referenced tests. Review of Educational Research, 56, 137–172.
- Berk, R. A. (1996). Standard setting: The next generation (where few psychometricians have gone before!). Applied Measurement in Education, 9, 215–235.
- Beuk, C. H. (1984). A method for reaching a compromise between absolute and relative standards in examinations. Journal of Educational Measurement, 21, 147–152.
- Board of Examiners, Uniform CPA Examination. (1996). 1996 annual report. New York: American Institute of Certified Public Accountants.
- Buckendahl, C., Plake, B. S., & Impara, J. C. (1999, April). Setting minimum passing scores on high-stakes assessments that combine selected response and constructed response formats. Paper presented at the annual meeting of the American Educational Research Association, Montreal, Quebec.

- Business Roundtable. (2000, September 13). New survey challenges extent of public backlash against state testing. Retrieved October 15, 2000, from <http://www.brtable.org/press.cfm/453>
- Chang, L. (1999). Judgmental item analysis of the Nedelsky and Angoff standard-setting methods. Applied Measurement in Education, 12, 151–165.
- Cizek, G. J. (1993). Reconsidering standards and criteria. Journal of Educational Measurement, 30, 93–106.
- Cizek, G. J. (1996a). Setting passing scores [An NCME instructional module]. Educational Measurement: Issues and Practice, 15(2), 20–31.
- Cizek, G. J. (1996b). Standard-setting guidelines. Educational Measurement: Issues and Practice, 15(1), 13–21, 12.
- Cizek, G. J. (2001). Conjectures on the rise and fall of standard setting: An introduction to context and practice. In G. J. Cizek (Ed.), Setting performance standards: Concepts, methods, and perspectives (pp. 3–17). Hillsdale, NJ: Erlbaum.
- Cizek, G. J., & Fitzgerald, S. M. (1996, April). A comparison of group and independent standard setting. Paper presented at the annual meeting of the American Educational Research Association, New York, NY.
- Cizek, G. J., & Husband, T. H. (1997, March). A Monte Carlo investigation of the contrasting groups standard setting method. Paper presented at the annual meeting of the American Educational Research Association, Chicago, IL.
- Cohen, A. S., Kane, M. R., & Crooks, T. J. (1999). A generalized examinee-centered method for setting standards on achievement tests. Applied Measurement in Education, 12, 327–366.
- Ebel, R. L. (1972). Essentials of educational measurement. Englewood Cliffs, NJ: Prentice-Hall.
- Geisinger, K. F. (1991). Using standard-setting data to establish cutoff scores. Educational Measurement: Issues and Practice, 10(2), 17–22.
- Giraud, G., Impara, J.C., & Buckendahl, C. (2000). Making the cut in school districts: Alternative methods for setting cut-scores. Educational Assessment, 6, 291–304.
- Glass, G. V. (1978). Standards and criteria. Journal of Educational Measurement, 15, 237–261.

- Goodwin, L. D. (1999). Relations between observed item difficulty levels and Angoff minimum passing levels for a group of borderline examinees. Applied Measurement in Education, 12, 13–28.
- Hambleton, R. K. (1998). Setting performance standards on achievement tests: Meeting the requirements of Title I. In L. Hansche (Ed.), Handbook for the development of performance standards: Meeting the requirements of Title I (pp. 87–114). Washington, DC: Council of Chief State School Officers.
- Hambleton, R. K., Brennan, R. L., Brown, W., Dodd, B., Forsyth, R. A., Mehrens, W. A., et al. (2000). A response to “Setting reasonable and useful performance standards” in the National Academy of Sciences’ Grading the nation’s report card. Educational Measurement: Issues and Practice, 19(2), 5–14.
- Hambleton, R. K., Jaeger, R. M., Plake, B. S., & Mills, C. N. (2000). Setting performance standards on complex educational assessments. Applied Psychological Measurement, 24, 355–366.
- Hambleton, R. K., Jaeger, R. M., Plake, B. S., & Mills, C. N. (in press). Handbook for setting performance standards. Washington, DC: Council of Chief State School Officers.
- Hambleton, R. K., & Plake, B. S. (1995). Using an extended Angoff procedure to set standards on complex performance assessments. Applied Measurement in Education, 8, 41–55.
- Hambleton, R. K., & Powell, S. (1983). A framework for viewing the process of standard-setting. Evaluation & the Health Profession, 6, 3–24.
- Impara, J. C., & Plake, B. S. (1998). Teachers’ ability to estimate item difficulty: A test of the assumptions in the Angoff standard setting method. Journal of Educational Measurement, 35, 69–81.
- Jaeger, R. M. (1982). An iterative structured judgment process for establishing standards on competency tests: Theory and application. Educational Evaluation and Policy Analysis, 4, 461–475.
- Jaeger, R. M. (1988). Use and effect of caution indices in detecting aberrant patterns of standard-setting judgments. Applied Measurement in Education, 1, 17–31
- Jaeger, R. M. (1989). Certification of student competence. In R. Linn (Ed.), Educational measurement (3rd ed., pp. 485–514). Englewood Cliffs, NJ: Prentice-Hall.
- Jaeger, R. M. (1991). Selection of judges for standard-setting. Educational Measurement, Issues and Practices, 10(2), 3–6, 10, 14.

- Jaeger, R. M. (1995). Setting performance standards through two-stage judgmental policy capturing. Applied Measurement in Education, 8, 15–40.
- Jaeger, R. M. & Mills, C. N. (2001). An integrated judgment procedure for setting standards on complex, large-scale assessments. In G. J. Cizek (Ed.), Standard setting: Concepts, methods, and perspectives (p. 313–338). Mahwah, NJ: Erlbaum.
- Kane, M. (1994). Validating the performance standards associated with passing scores. Review of Educational Research, 64, 425–461.
- Kane, M. (1995). Examinee-centered vs. task-centered standard setting. In Proceedings of the joint conference on standard setting for large scale assessments of the National Assessment Governing Board (NAGB) and the National Center for Educational Statistics (NCES), Volume II (pp. 119–141). Washington, DC: U. S. Government Printing Office.
- Kane, M. (1998). Choosing between examinee-centered and test-centered standard-setting methods. Educational Assessment, 5, 129–145.
- Kane, M. (2001). So much remains the same: Conception and status of validation in setting standards. In G. Cizek (Ed.), Standard setting: Concepts, methods, and perspectives (pp. 53–88). Mahwah, NJ: Erlbaum.
- Kane, M., Crooks, T., & Cohen, A. (1997, March). Justifying the passing scores for licensure and certification tests. Paper presented at the annual meeting of the American Educational Research Association, Chicago, IL.
- Kingston, N. M., Kahl, S. R., Sweeney, K., & Bay, L. (2001). Setting performance standards using the body of work method. In G. J. Cizek (Ed.), Standard setting: Concepts, methods, and perspectives (pp. 219–248). Mahwah, NJ: Erlbaum.
- Lewis, D. M., Mitzel, H. C., & Green, D. R. (1996, June). Standard setting: A bookmark approach. Paper presented at the meeting of the Council of Chief State School Officers, Phoenix, AZ.
- Linn, R. L. (1998). Validating inferences from National Assessment of Educational Progress achievement-level reporting. Applied Measurement in Education, 11, 23–47.
- Livingston, S. A., & Zieky, M. J. (1982). Passing scores: A manual for setting standards of performance on educational and occupational tests. Princeton, NJ: Educational Testing Service.
- Livingston, S. A., & Zieky, M. J. (1989). A comparative study of standard-setting methods. Applied Measurement in Education, 2, 121–141.

- Lord, F. M., & Novick, M. R. (1968). Statistical theories of mental test scores. Reading, MA: Addison-Wesley.
- Massachusetts Department of Education. (2000). Massachusetts Comprehensive Assessment System performance level definitions. Retrieved October 15, 2000, from <http://www.doe.mass.edu/mcas/mcaspld.html>
- Meara, K. C. (2000). Validity issues in standard setting. Unpublished doctoral dissertation, University of Massachusetts, Amherst.
- Mehrens, W. A. (1995). Methodological issues in standard setting for educational exams. In Proceedings of the joint conference on standard setting for large scale assessments of the National Assessment Governing Board (NAGB) and the National Center for Educational Statistics (NCES), Volume II (pp. 221–263). Washington, DC: U. S. Government Printing Office.
- Melican, G. J., Mills, C. N., & Plake, B. S. (1989). Accuracy of item performance predictions based on the Nedelsky standard setting method. Educational and Psychological Measurement, 49, 467–478.
- Mills, C. N. (1995). Establishing passing standards. In J. C. Impara (Ed.), Licensure testing: Purposes, procedures, and practices (pp. 219–252). Lincoln, NE: Buros Institute of Mental Measurements.
- Mills, C. N., Hambleton, R. K., Biskin, B., Kobrin, J., Evans, J., & Pfeffer, M. (2000, January). A comparison of two standard setting methods for the Uniform CPA Examination: Technical report. Unpublished manuscript.
- Mitzel, H. C., Lewis, D. M., Patz, R. J., & Green, D. R. (2001). The bookmark procedure: Psychological perspectives. In G. J. Cizek (Ed.), Standard setting: Concepts, methods, and perspectives (pp. 249–281). Mahwah, NJ: Erlbaum.
- Morse, J. (2000, September 11). Does Texas make the grade? Time, 156(11), 50–54.
- Nedelsky, L. (1954). Absolute grading standards for objective tests. Educational and Psychological Measurement, 14, 3–19.
- Norcini, J. J. (1994). Research on standards for professional licensure and certification examinations. Evaluation and the Health Professions, 17, 160–177.
- Norcini, J. J., & Shea, J. A. (1997). The credibility and comparability of standards. Applied Measurement in Education, 10, 39–59.
- Pellegrino, J. W., Jones, L. R., & Mitchell, K. J. (1999). Grading the nation's report card: Evaluating NAEP and transforming the assessment of educational progress. Washington, DC: National Academy Press.

- Pitoniak, M. J., & Sireci, S. G. (1999). A literature review of contemporary standard-setting methods appropriate for computerized-adaptive tests (Laboratory of Psychometric and Evaluative Methods Research Report No. 369). Amherst: University of Massachusetts, School of Education.
- Pitoniak, M. J., Sireci, S. G., & Luecht, R. M. (in press). A multitrait-multimethod validity investigation of scores from a professional licensure examination. Educational and Psychological Measurement.
- Plake, B. S. (1995). An integration and reprise: What we think we have learned. Applied Measurement in Education, 8, 85–92.
- Plake, B. S. (1997). Criteria for evaluating the quality of a judgmental standard setting procedure: What information should be reported? Unpublished manuscript.
- Plake, B. S. (1998). Setting performance standards for professional licensure and certification. Applied Measurement in Education, 11, 65–80.
- Plake, B. S., & Hambleton, R. K. (2000). A standard-setting method designed for complex performance assessments: Categorical assignments of student work. Educational Assessment, 6, 197–215.
- Plake, B. S., & Hambleton, R. K. (2001). The analytic judgment method for setting standards on complex performance assessments. In G. J. Cizek (Ed.), Standard setting: Concepts, methods, and perspectives (pp. 283–312). Mahwah, NJ: Erlbaum.
- Plake, B. S., Hambleton, R. K., & Jaeger, R. M. (1997). A new standard setting method for performance assessments: The dominant profile judgment method and some field-test results. Educational and Psychological Measurement, 57, 400–411.
- Plake, B. S., Impara, J. C., & Irwin, P. (1999, April). Validation of Angoff-based predictions of item performance. Paper presented at the annual meeting of the American Educational Research Association, Montreal, Quebec.
- Putnam, S. E., Pence, P., & Jaeger, R. M. (1995). A multi-stage dominant profile method for setting standards on complex performance assessments. Applied Measurement in Education, 8, 57–83.
- Reckase, M. D., & Bay, L. (1999, April). Comparing two methods for collecting test-based judgments. Paper presented at the annual meeting of the National Council on Measurement in Education, Montreal, Quebec.
- Shepard, L. (1980). Standard setting issues and methods. Applied Psychological Measurement, 4, 447–467.

- Shepard, L. A. (1995). Implications for standard setting of the National Academy of Education evaluation of the National Assessment of Educational Progress achievement levels. In Proceedings of the joint conference on standard setting for large scale assessments of the National Assessment Governing Board (NAGB) and the National Center for Educational Statistics (NCES), Volume II (pp. 143–160). Washington, DC: U. S. Government Printing Office.
- Shepard, L., Glaser, R., Linn, R., & Bohmstedt, G. (1993). Setting performance standards for student achievement. Stanford, CA: National Academy of Education.
- Sireci, S. G. (1995, August). Using cluster analysis to solve the problem of standard setting. Paper presented at the annual meeting of the American Psychological Association, New York, NY.
- Sireci, S. G. (2001). Standard setting using cluster analysis. In G. J. Cizek (Ed.), Standard setting: Concepts, methods, and perspectives (pp. 339–354). Mahwah, NJ: Erlbaum.
- Sireci, S. G., & Biskin, B. J. (1992). Measurement practices in national licensing examination programs: A survey. CLEAR Exam Review, 3(1), 21–25.
- Sireci, S. G., & Green, P. C. (2000). Legal and psychometric criteria for evaluating teacher certification tests. Educational Measurement: Issues and Practice, 19(1), 22–31, 34.
- Sireci, S. G., Hambleton, R. K., Huff, K. L., & Jodoin, M. G. (2000). Setting and validating standards on Microsoft Certified Professional examinations (Laboratory of Psychometric and Evaluative Research Report No. 395). Amherst: University of Massachusetts, School of Education.
- Sireci, S. G., Pitoniak, M. J., Meara, K. C., & Hambleton, R. K. (2000). A review of standard setting methods applicable to the Advanced Placement examination program (Laboratory of Psychometric and Evaluative Research Report No. 375). Amherst: University of Massachusetts, School of Education.
- Sireci, S. G., Rizavi, S., Dillingham, A., & Rodriguez, G. (1999). Setting performance standards on the ACCUPLACER Elementary Algebra Test (Laboratory of Psychometric and Evaluative Research Report No. 368). Amherst: University of Massachusetts, School of Education.
- Sireci, S. G., Robin, F., & Patelis, T. (1999). Using cluster analysis to facilitate standard setting. Applied Measurement in Education, 12, 301–325.

- State of Wisconsin, Department of Public Instruction. (1997). Final summary report of the proficiency score standards for the Wisconsin Student Assessment System (WSAS) Knowledge and Concept Examinations for elementary, middle, and high school at grades 4, 8, and 10. Madison, WI: Office of Educational Accountability.
- Subkoviak, M. J., Kane, M. T., & Duncan, P.H. (1999, April). A comparative study of the Angoff and Nedelsky methods: Implications for validity. Paper presented at the annual meeting of the National Council on Measurement in Education, Montreal, Quebec.
- United States General Accounting Office. (1993). Educational achievement standards: NAGB's approach yields misleading interpretations (GAO/PEMD Publication No. 93-12). Washington, DC: Author.
- van der Linden, W. J. (1982). A latent trait method for determining intrajudge consistency in the Angoff and Nedelsky techniques of standard setting. Journal of Educational Measurement, 19, 295-308.
- van der Linden, W. J. (1995). A conceptual analysis of standard setting in large-scale assessments. In Proceedings of the joint conference on standard setting for large scale assessments of the National Assessment Governing Board (NAGB) and the National Center for Educational Statistics (NCES), Volume II (pp. 97-117). Washington, DC: U. S. Government Printing Office.

