THE EFFECTS OF INTRAJUDGE CONSISTENCY FEEDBACK IN AN ANGOFF STANDARD-SETTING PROCEDURE

A DISSERTATION SUBMITTED TO THE GRADUATE DIVISION OF THE UNIVERSITY OF HAWAI‘I AT MĀNOA IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF

DOCTOR OF PHILOSOPHY

IN

EDUCATIONAL PSYCHOLOGY

AUGUST 2013

By

George M. Harrison

Dissertation Committee:

Paul R. Brandon, Chairperson
James D. Brown
Ronald H. Heck
Seongah Im
Nicole R. Lewis

Keywords: Standard setting, modified Angoff method, cutscore, intrajudge consistency, interjudge consistency, multilevel modeling, generalizability theory
ACKNOWLEDGEMENTS

I would like to thank my dissertation committee members for their expertise, time, and moral support. Their enthusiasm and thoughtful questions have helped me throughout this dissertation process.

Special thanks goes to Professor Paul Brandon, my chair and mentor. I am particularly grateful for his guidance, feedback, and thoughtful review of my work.

I would also like to thank all of those who participated as judges in this study, particularly those who provided me with feedback in the pilot-testing sessions.

Thanks also goes to my colleagues at the Curriculum Research & Development Group and the University Laboratory School, whose support and encouragement helped me along the way.

Finally, I would like to acknowledge my wife for her patience, support, and understanding. Thank you for reminding me what is truly important in life.
ABSTRACT

Agencies establishing performance levels on tests utilize standard-setting procedures to derive cutscores for making classificatory decisions about examinees. The credibility of standard-setting cutscores depends, in part, on two sources of internal validity evidence: intrajudge and interjudge consistency. Feedback to improve intrajudge consistency has been routinely suggested, but scarcely experimentally tested. This dissertation investigates the effect of item-level intrajudge-consistency feedback on changes in intrajudge and interjudge consistency. In this study, participants with secondary- or post-secondary teaching experience served as Angoff judges, making three rounds of judgments about the probability of success of conceptualized barely proficient examinees (BPEs) on 50 vocabulary-test items. Using a randomized experimental design, I assigned participants to either a treatment \( (n = 18) \) or control \( (n = 18) \) group and facilitated 23 standard-setting sessions. Treatment-group judges received item-level intrajudge-consistency feedback; control-group judges performed an alternative between-round task. Using a multilevel-model-for-change framework, I compared the two groups in their round-to-round changes in consistency indexes. Using generalizability theory, I investigated the changes in interjudge consistency and estimated the minimum number of judges needed to achieve a degree of precision specified in previous research. Results from the multilevel analysis indicated that improvements in intrajudge consistency were significantly greater for the treatment group \( (p < .001) \). Generalizability-theory results provided evidence of improved interjudge consistency: From Round 1 to 3, unexplained variance decreased from 36% to 23%, dependability improved from .94 to .96, and estimates of the fixed-item standard error of the cutscore decreased from 1.49 to 1.38.
Decision-study results revealed diminishing returns in precision after about 10 judges. The findings suggest that item-level intrajudge-consistency feedback improves judges’ accuracy in providing ratings that are consistent with their individual conceptualizations of the BPE. The feedback likely improves interjudge consistency by reducing variability attributed to idiosyncratic item ratings among judges. Decision-study results suggest that not only are about 10 judges sufficient for similar Angoff procedures, but also that feedback provides a benefit equivalent to hiring 2 judges. These findings contribute to the growing body of research on standard-setting feedback and provide empirical evidence for practitioners planning Angoff procedures.
CONTENTS

ACKNOWLEDGEMENTS .............................................................................................................. ii
ABSTRACT .................................................................................................................................. iii
TABLES ......................................................................................................................................... ix
FIGURES ....................................................................................................................................... xi

CHAPTER 1: INTRODUCTION ................................................................................................ 1
  Background ............................................................................................................................. 1
  Statement of the Purpose ...................................................................................................... 5
  Definition of Terms ............................................................................................................... 7

CHAPTER 2: REVIEW OF THE LITERATURE ..................................................................... 8
  Credibility of Angoff-method Procedures ........................................................................ 8
    The Angoff method .......................................................................................................... 8
    Criticism of the Angoff method ...................................................................................... 9
    Credibility of cutscores .................................................................................................. 9
  Measures of Intrajudge Consistency ............................................................................. 13
    Between-round measures .............................................................................................. 14
    Within-round measures ............................................................................................... 15
  Measures of Interjudge Inconsistency under a Generalizability-theory Framework ... 23
    Variance components ................................................................................................. 25
    Estimating the minimum number of judges needed to meet a precision criterion.... 29
  Using Intrajudge and Interjudge Consistency Data to Evaluate Cutscore Credibility. 29
  Item-level Feedback ......................................................................................................... 31
    Conditional \( p \)-value feedback .................................................................................. 32
    Intrajudge-consistency feedback............................................................................... 34
Research Questions ................................................................. 41

CHAPTER 3: METHODS ............................................................... 45

Participants ............................................................................. 45

Materials .................................................................................. 48

Test materials .......................................................................... 48

Background questionnaire .................................................... 53

Random-group assignment materials .................................. 54

Facilitator protocol checklist .................................................. 55

Computers and software ....................................................... 55

Session observation log ......................................................... 56

Orientation sheets .................................................................... 56

Standard-setting entry forms ............................................... 57

Post-session questionnaires .................................................. 64

Data Collection ........................................................................ 65

Pilot testing ............................................................................ 65

Participant recruitment .......................................................... 67

Background questionnaire data collection ............................ 70

Random group assignment .................................................... 71

Standard-setting sessions ...................................................... 76

Preparation of the Data for Analyses ................................. 81

Raw data ................................................................................ 81

Participant-background covariate variables ...................... 81

Analysis for Research Question 1 ....................................... 82

Statistical assumptions ......................................................... 82
Data screening ........................................................................................................... 84
Model selection ......................................................................................................... 85
Analysis for Research Questions 2 and 3 ................................................................. 93
CHAPTER 4: RESULTS ...................................................................................................95
Intrajudge Consistency Indexes ................................................................................... 95
Descriptive statistics ................................................................................................. 95
Reliability of the consistency indexes ....................................................................... 97
Normality and outlier detection ................................................................................ 98
Participant-background Covariates ............................................................................ 100
Descriptive statistics ............................................................................................... 100
Reliability of the covariates .................................................................................... 101
Correlations among covariates ................................................................................ 101
Correlations between each covariate and each round’s consistency index .......... 101
Model Comparisons for Predicting Intrajudge Consistency Indexes ......................... 102
Unconditional means model.................................................................................... 103
Unconditional growth model .................................................................................. 103
Experimental group as predictor ............................................................................. 105
Participant-background covariate models .............................................................. 107
Results of the first stage of model comparisons ..................................................... 107
Covariance structure comparisons .......................................................................... 108
Intercepts- and intercepts-and-slopes-as-outcomes models..................................... 109
Results of the second stage of model comparisons ................................................. 111
Generalizability-theory Results for Examining Interjudge Consistency ................... 114
Descriptive statistics of the cutscores ..................................................................... 115
TABLES

Table 2.1 Consistency Indexes Reported in Empirical Studies ........................................ 19
Table 2.2 Percent of Variance and Dependability Indexes in Pre-feedback Conditions in Item-by-judge Angoff Studies .................................................. 26
Table 2.3 Studies Reporting Variance Components before and after Conditional p-value Feedback ................................................................................................ 34
Table 3.1 Number of Participants and Years of Teaching Experience in Relevant Contexts ................................................................. 46
Table 3.2 Number of Participants in Control and Treatment Groups with Relevant Teaching Experience ............................................................... 47
Table 3.3 Number of Participants with Experience Teaching Each Secondary Level Grade ............................................................ 47
Table 3.4 Number of Participants with Experience Teaching Each College Discipline ........................................................................ 48
Table 3.5 Criteria Used in Selecting Items for Use in the Standard-setting Procedures .......................................................... 50
Table 3.6 Descriptive Statistics of the Empirical Item Difficulty Estimates ..................... 53
Table 3.7 Normality of the Empirical Item Difficulty Estimates .................................. 53
Table 3.8 Features of the Instruments Investigated During Pilot Testing ....................... 66
Table 3.9 Recommendations on the Number of Judges on an Angoff Standard-setting Panel ........................................................................ 68
Table 3.10 Practice Session Control and Treatment Activities ....................................... 79
Table 3.11 Operational Session Control and Treatment Activities ............................. 80
Table 3.12 Specifications of Models in the First Stage of Model Comparisons .............. 88
Table 4.1 Descriptive Statistics of the Consistency Indexes per Round ......................... 96
Table 4.2 Univariate Tests of Normality of Consistency Indexes in Each Round of Rating .................................................................................... 98
Table 4.3 Descriptive Statistics of the Participant-background Covariate Scores ....... 101
Table 4.4 Correlations between Participant-background Covariates and Consistency Indexes .............................................................................................................. 102

Table 4.5 Parameter Estimates in the First Stage of Model Comparisons ............... 104

Table 4.6 Fit Indexes of the Logarithmic Fixed-effects Model with Various Covariance Matrix Structures .............................................................................. 108

Table 4.7 Comparison of Covariance Structure Models ............................................. 109

Table 4.8 Fit Indexes of Converging Random-effects Models and the Best Fixed-effects Model ................................................................................................ 110

Table 4.9 Parameter Estimates of the Optimal Model ................................................. 111

Table 4.10 Normality of the Residuals of the Optimal Model .................................... 113

Table 4.11 Descriptive Statistics of the Cutscores per Group per Round ..................... 116

Table 4.12 Treatment Group’s Estimated Variance Components in Each of the Three Rounds ........................................................................................................ 117

Table 4.13 Control Group’s Estimated Variance Components in Each of the Three Rounds ........................................................................................................ 118

Table 4.14 Dependability Indexes and Estimates of the Standard Error of the Cutscore ........................................................................................................ 119

Table 4.15 D Study of the Round 1 Ratings ................................................................. 121

Table 4.16 D Study of the Round 3 Ratings ................................................................. 122

Table H.1 Group Assignment Probabilities in the Secondary-experience Stratum ...... 177

Table H.2 Group Assignment Probabilities in the Other Stratum ............................. 178
FIGURES

Figure 2.1 A Hypothetical Judge’s Observed and Expected Angoff Rating .................. 17

Figure 4.1 Plot of the Mean Observed Consistency Index per Round ...................... 97

Figure 4.2 Normal Probability Plots of Consistency Indexes in Each Round per Experimental Group ................................................................. 99

Figure 4.3 Quantile-quantile Plot of the Residuals of the Optimal Model ............ 113

Figure 4.4 Plot of the Fixed-effects Logarithmic Model with Toeplitz Covariance Structure .......................................................................................... 115

Figure 4.5 D-study Plot of the Round 3 Standard Errors of the Cutscore with Varying Numbers of Judges ................................................................. 123

Figure 4.6 D-study Plot of the Round 3 Dependability-indexes with Varying Numbers of Judges ................................................................. 123
CHAPTER 1

INTRODUCTION

Background

Standard setting is an activity in which judges on a panel decide where to place one or more cutscores on a test in order to classify or make decisions about examinees (Cizek, 1996, 2001; Cizek & Bunch, 2007; Hambleton & Pitoniak, 2006). Judges, also called panelists or participants, are typically professionals, such as teachers or subject-matter experts, recruited by a standard-setting agency or other organization to participate in activities that result in the panel’s recommended cutscore. A cutscore is the test score that is used for classifying examinees to determine whether they have or do not have enough knowledge, skills, or abilities to be considered proficient (or to be considered prepared, ready, advanced, and so forth). This degree of proficiency, which the cutscore operationalizes, is a performance standard. Unlike content standards, which are commonly used by agencies for defining what should be learned, performance standards represent how much examinees are expected to know.

Nearly every professional certification agency, state department of education, and many school districts participate in the setting of standards for their assessments (Hambleton & Pitoniak, 2006). Since the late 1970s, researchers have devoted a great deal of attention to better understanding the details of the standard-setting process (Zieky, 2001). However, even after three decades of research into standard setting, many policymakers, researchers, and educators are still uneasy about current standard-setting methods (Hambleton & Pitoniak, 2006). Experts (e.g., Cizek & Bunch, 2007) still make calls for more studies investigating the details in standard-setting procedures, particularly
in ways that judges use information in deciding where to place a cutscore (Cizek & Bunch, 2007; Reckase, 2001).

Over the past 40 years, the most commonly used standard-setting method in education has been the Angoff method (Angoff, 1971) and its modifications (Plake & Cizek, 2012). In the Angoff method, judges are first asked to conceptualize a hypothetical examinee who has barely enough knowledge, skill, or ability to be classified as meeting the performance standard. Judges are then tasked with estimating the probability of success of this type of examinee on each dichotomously-scored item on a test. Multiple rounds of rating of the same set of items are usually conducted to ensure judges understand the task and accurately translate their conceptualizations into estimates of probability of success. Multiple judges, making up a panel, are also included in the process to ensure that the resulting cutscore is reasonable and not based on the judgment of a single individual. Typically, judges come from different stakeholder groups, including people who are familiar with the examinee population and individuals with a stake in the outcome of the use of the cutscore (Hambleton & Pitoniak, 2006; Loomis, 2012; Raymond & Reid, 2001).

Although it is well established that different standard-setting methods typically yield different cutscores (Jaeger, 1989), an understudied component of standard setting which may equally affect cutscore estimation is the feedback that standard-setting judges review between rounds in standard-setting procedures (Reckase, 2001). Feedback is often included in standard-setting procedures with the intention of improving the procedural and internal validity of the process (Kane, 1994). Experts (Reckase & Chen, 2012) assert that “feedback helps panelists learn to make more accurate judgments during the standard
setting process so that their judgments will improve in quality over rounds” (p. 152). Feedback activities are assumed to minimize errors that might occur when judges misunderstand the procedure or inaccurately apply their judgments in the task (Loomis, 2012; Reckase, 2001).

Many types of feedback (which I briefly describe in Chapter 2) have been used in the Angoff method (Cizek & Bunch, 2007; Reckase, 2001). Agencies planning standard-setting procedures must be selective in determining which feedback activities to include because, according to research based on teachers’ self-reports (e.g., Loomis, 2000), having judges review too many kinds of feedback can cause them to ignore or misunderstand information provided to them (Reckase & Chen, 2012). Investigations into feedback effects (e.g., Busch & Jaeger, 1990; Clauser, Harik, et al., 2009; Clauser, Mee, Baldwin, Margolis, & Dillon, 2009; Loomis, 2000; Norcini, Shea, & Kanya, 1988; Yang, 2000) have mostly been conducted in operational settings, in which the primary purpose was for setting a cutscore rather than manipulating feedback activities to examine their effects. As pointed out by Reckase (2001), operational standard-setting procedures often employ multiple types of feedback, making it difficult to isolate the effect of a specific type of feedback (cf., Clauser, Harik, et al., 2009; Clauser, Mee, et al., 2009). According to scholars (Hambleton, Pitoniak, & Copella, 2012; Reckase, 2001; Reckase & Chen, 2012), more research is needed on the effects of feedback. To the best of my knowledge, there have been no experiments of the causal effects of specific kinds of feedback comparing judges receiving a treatment with those receiving alternative between-round activities. Studies revealing these causal effects will facilitate good decision-making among agencies planning standard-setting procedures. In this study, I examine one type
of feedback, intrajudge consistency (van der Linden, 1982), and its effect on the error variability within and between judges.

Error variability in the judgment process can arise from within-judge (intrajudge) inconsistency. From a psychometric perspective (Camilli, Cizek, & Lugg, 2001; Nichols, Twing, Mueller, & O’Malley, 2010; Reckase, 1998, 2006), Angoff judges are tasked with translating their representations of the performance standard, the psychological construct, into a quantifiable judgment about success on each item, which is the operationalization of this construct. A judge’s observed Angoff rating on an item represents one of many possible ratings for that item (Reckase, 2006). Van der Linden’s (1982) intrajudge-consistency index provides a measure of this within-judge variability that is independent of other judges on the panel. This measure uses what is known about each item’s empirical functioning, based on previous administrations of the item to examinees, and what is assumed to be the judge’s true intended cutscore, based on the judgments across the entire set of items.

Error variability can also be examined as between-judge (interjudge) inconsistencies among judges in a panel. This error comprises two components: discrepancies in judges’ cutscores and discrepancies in judges’ item-level Angoff ratings. Judges may differ in their recommended cutscores because they differ in their representations of the performance-standard construct (because of their different values or beliefs about how good is good enough). This variability contributes to estimates of the standard error of the cutscore (Brennan & Lockwood, 1980; Hurtz & Hertz, 1999), which agencies use for evaluating the credibility of the panel’s recommended cutscore (Loomis, 2012). Judges can also differ at the item level. That is, after cutscore variability has been taken into
account, judges may differ in their translations of their representations of the construct into item-level ratings. This can be construed as idiosyncratic judgments conflated with unexplained variability. This is similar to intrajudge inconsistency but with discrepancies being in relation to other judges’ ratings rather than in relation to what is known about empirical item functioning. Generalizability-theory (Cronbach, Rajaratnam, & Gleser, 1963) provides a framework for measuring error variability at both the item-level (item-by-judge component) and overall cutscore (judge component) level (Brennan, 1995; Brennan & Lockwood, 1980; Raymond & Reid, 2001).

**Statement of the Purpose**

The purpose of this study is to investigate the effects of item-level intrajudge-consistency feedback on Angoff standard-setting judges’ intrajudge and interjudge consistency. Intrajudge consistency is operationalized at the judge level using van der Linden’s (1982) consistency index. Interjudge consistency is operationalized as judge and item-by-judge variance components estimated from a generalizability-theory framework. Measures of intrajudge and interjudge consistency provide information for evaluating the credibility of the standard-setting procedure (Hambleton et al., 2012).

This study differs from previous work in feedback in standard setting in two ways: (a) intrajudge-consistency feedback is isolated from other between-round feedback, such as impact data (the percentage of students above the cutscore), discussion, and presentation of other panelists’ ratings, and (b) judges receiving the feedback are compared with a control group of judges participating in a similar activity but without the feedback.
I address three research questions, which I more fully explain in Chapter 2. The first investigates whether item-level intrajudge-consistency feedback has an effect on judge-level intrajudge consistency. More accurately, I look at whether there is a difference between two groups of judges’ changes in intrajudge consistency over the course of three standard-setting rounds when the feedback provided to judges between rounds differs between the two groups.

The second research question addresses the degree to which item-level intrajudge-consistency feedback improves interjudge consistency. Specifically, I employ a generalizability-theory framework and examine the two groups of judges’ changes in variance components, dependability indexes, and standard errors of the cutscore.

The final research question is a continuation of Research Question 2, and asks how many judges would be required to achieve an acceptable degree of precision. Specifically, I use criteria described in Hurtz and Hertz (1999) and Raymond and Reid (2001) and conduct a decision study to estimate the minimum number of judges required to achieve these criteria. I also examine the treatment and control groups’ Round 1 to Round 3 changes in these estimates to illustrate the effect of intrajudge-consistency feedback on interjudge consistency in terms of the number of judges needed.

To conduct the study, I employed a randomized-experiment design (Shadish, Cook, & Campbell, 2002) and facilitated a three-hour standard-setting session with each participant. I recruited 36 secondary and post-secondary teachers to serve as judges and assigned half to the treatment \((n = 18)\) and half to the control \((n = 18)\) group. Treatment-group judges received item-level intrajudge-consistency feedback while control-group
judges received an alternative between-round activity that required approximately the same amount of time and attention.

This study contributes to the research base on feedback in standard-setting procedures. It also provides practical evidence useful to standard-setting agencies seeking to make informed decisions about which types of feedback to include in their standard-setting procedures.

**Definition of Terms**

When practitioners and researchers use the term *Angoff method*, they are often referring to the family of methods that employ multiple-round, modified, Angoff method procedures (Cizek & Bunch, 2007). I follow this convention, using the term *Angoff method* to refer to both the procedures Angoff (1971) described (in his footnote) and to any procedure (including modified Angoff procedures) in which judges iteratively rate each item by participating in two or more rounds of rating. I use the terms *Angoff rating*, *rating*, and *probability judgment* to refer to the rating that a judge or panel of judges provides on an item to estimate the probability that the hypothetical examinee will correctly respond to a dichotomously scored item.

Throughout the literature, the terms *judge*, *panelist*, *rater*, and *participant* have been used to refer to an individual providing Angoff ratings on items; I also use these terms interchangeably to avoid awkward sentence construction. I employ the term *barely proficient examinee* (BPE) to describe the hypothetical examinee with just enough knowledge, skills, or abilities to be categorized as proficient; other authors (e.g., Hambleton & Pitoniak, 2006; Hurtz & Auerbach, 2003) label the BPE *minimally competent, minimally qualified*, or *borderline*. 
CHAPTER 2

REVIEW OF THE LITERATURE

To frame my study in the context of the existing literature, I introduce the Angoff method and discuss a concern that has been raised among critics of the method—namely, that the credibility of an Angoff procedure’s results depends in part on evidence of how consistent judges are in applying their judgments across items. I then describe the methods that have been used to measure intrajudge and interjudge consistency and address what is known about the effects of item-level feedback on these two types of consistency. I end this chapter with a discussion of the study’s research questions.

Credibility of Angoff-method Procedures

The Angoff method. The most commonly used and well-researched standard-setting approach is the modified Angoff method (Angoff, 1971) and its variations (Cizek & Bunch, 2007; Hurtz & Auerbach, 2003; Plake & Cizek, 2012). In its most basic form, which Angoff (1971) described in a footnote as an adaptation to a procedure developed by Ledyard Tucker, each judge on a panel rates each dichotomously-scored item on a test according to the probability that a barely proficient examinee will be able to answer the item correctly. Often, the judge is asked to estimate the percentage of a hypothetical group of examinees—100 BPEs, for instance—who would respond correctly. The sum of a judge’s item probability estimates is that judge’s recommended cutscore. The panel’s recommended cutscore is calculated as the mean of all the judges’ recommended cutscores.

The modifications of the method that Angoff (1971) described involves having judges carry out two or three iterations of the rating procedure, with information provided
to the judges between rounds. Variations in the modified Angoff method are
categorized by the types of information judges receive and review between iterations
(Cizek & Bunch, 2007; Hambleton & Pitoniak, 2006; Hurtz & Auerbach, 2003; Plake &
Cizek, 2012), one of which can be feedback given to judges about the consistency in their
Angoff ratings. Typically, the recommended cutscore is determined in the final round and
used by the board or agency that ultimately decides where the cutscore will be placed
(Cizek & Bunch, 2007).

**Criticism of the Angoff method.** A criticism of the Angoff method has been that
judges have difficulty making probability judgments while simultaneously keeping a
conceptualization of the BPE in mind (Impara & Plake, 1998; Shepard, 1995; Shepard,
report (Shepard et al., 1993) characterized the Angoff method as “fundamentally flawed”
(p. xxiv), claiming that panelists were incapable of making consistent and reasonable
judgments. Many of the methods and findings of the NAE report, however, have been
found to be untenable by standard-setting scholars (Cizek, 1993; Cizek & Bunch, 2007;
Hambleton, 2001; Hambleton et al., 2000; Hambleton & Pitoniak, 2006; Kane, 1995;
Loomis & Bourque, 2001). Concern with judges’ ability to provide consistent and
accurate probability judgments, nonetheless, persists, as is evident by the large amount of
research investigating Angoff judges’ consistency (see Brandon, 2004).

**Credibility of cutscores.** Because there is no true cutscore that exists in nature,
standard-setting is an inherently judgmental process (Camilli et al., 2001; Hambleton,
scholar, Glass (1978), went so far as to characterize the setting of cutscores as completely
arbitrary and recommended that the practice be abandoned altogether. Scholars
(Hambleton, 1978; Popham, 1978; Scriven, 1978) responding in the same Journal of
Educational Measurement issue, however, rebutted Glass’s claim, making the case that
with well-informed and careful judgment, credible cutscores, while still arbitrary, can be
arrived at. It is now generally accepted that what must be avoided are capricious
judgments (Cizek & Bunch, 2007; Hambleton & Pitoniak, 2006; Popham, 1978; Scriven,
1978), which can result from judges’ inconsistent translations of conceptualized BPE
ability into item ratings (van der Linden, 1982), carelessness, or lack of concern in how
the cutscore will be used (Camilli et al., 2001).

Credibility in a recommended cutscore depends in part on documentation that the
standard-setting procedures have been carried out appropriately, for procedural evidence
of validity, and in part on evidence that judges’ ratings are internally consistent (Kane,
1994). Important for establishing procedural validity is evidence that steps have been
taken to ensure that judges accurately translate their conceptualizations of the BPE into
item ratings (Kane, 1994). As prescribed by the Standards for Educational and
Psychological Testing (American Educational Research Association, American
Psychological Association, & National Council on Measurement in Education [AERA,
APA, & NCME], 1999), which I hereafter refer to as the Standards, “the judgmental
process should be designed so that judges can bring their knowledge and experience to
bear in a reasonable way” (p. 60); furthermore, “the procedures used to elicit such
judgments should result in reasonable, defensible standards that accurately reflect the
judges’ values and intentions” (p. 60). In the modified Angoff method, an important
procedural component is to have judges participate in between-round activities that
provide feedback about the judgment process (Reckase, 2001). Evidence that item ratings are consistent with judges’ knowledge, experience, and values assuages concerns that cutscores may have been made capriciously.

**The purpose of between-round feedback.** Participating in iterative rounds of rating allows judges to change their ratings after reviewing feedback between rounds, a process intended to improve probability judgments and, in turn, ensure credibility in the recommended cutscore. Feedback serves to inform judges of the accuracy of their understandings of the rating procedures and of their conceptualizations of the BPE’s abilities in relation to existing empirical data (Reckase, 2001; Reckase & Chen, 2012). According to Hambleton (2001), “the purpose of the discussion and feedback are to provide opportunity for panelists to reconsider their initial ratings and to identify errors or any misconceptions or misunderstandings that may be present” (p. 101). Thus, one of the key purposes of providing feedback is to reduce errors in panelists’ judgments (Hambleton & Pitoniak, 2006; Hurtz & Auerbach, 2003; Kane, 1994).

**Errors in Angoff ratings.** In Angoff procedures, error can be examined from two different measurement perspectives: (a) inconsistencies in a judge’s translations of his or her conceptualization of the BPE’s ability into item-level ratings and (b) discrepancies among judges in their item-level ratings and in their resulting cutscores. The first kind of error is intrajudge inconsistency and depends on how consistent or stable an individual judge is in his or her item-level ratings regardless of the judgments made by other panelists. This is important when evaluating whether individual judgments were made capriciously (van der Linden, 1982). The second is interjudge inconsistency and depends on the degree of consistency or convergence among the judges. This is important when
examining the precision of the cutscore (Brennan & Lockwood, 1980; Hurtz & Hertz, 1999; Jaeger, 1991; Kane, 1994) and has practical implications for agencies when evaluating or planning a standard-setting procedure (Hambleton & Pitoniak, 2006; Loomis, 2012) and when ultimately deciding where to place the operational cutscore based on the judges’ recommended cutscores (Geisinger & McCormick, 2010).

When a broad range of stakeholders makes up the panel of judges, as is often the case in education standard setting, judges may vary in their values or beliefs about what constitutes barely proficient performance, resulting in low precision (Cizek, 1996; Hambleton & Pitoniak, 2006; Kane, 1994, 2001; Loomis, 2012). In some circumstances, however, this variability may not necessarily be a threat to validity, because the inclusion of multiple stakeholders with differing values and expertise may be crucial for establishing the credibility of the cutscore (Cizek, 1996; Kane, 1994; Loomis, 2012). Nonetheless, the degree of variability among judges will be important when determining the minimum number of judges needed to achieve a reasonable degree of precision in the cutscore (Brennan, 1995; Hurtz & Hertz, 1999; Kane, 1994; Raymond & Reid, 2001).

In this study, I examine both types of error, intrajudge and interjudge inconsistency, and the effect that one type of feedback (item-level intrajudge-consistency feedback) has on them. Provision of feedback strengthens the procedural validity of the standard-setting process in that steps are taken to reduce error (Kane, 1994). If error is indeed minimized, feedback contributes to the internal validity argument of the resulting cutscores (Hambleton & Pitoniak, 2006; Hambleton et al., 2012).

**Types of feedback in Angoff-method procedures.** Feedback in the Angoff method occurs between rounds and can include (a) impact data (the percent of examinees that
would successfully meet the performance standard recommended in the previous round); (b) interjudge consensus or rater-location data (informing judges of the similarity of their ratings with that of the other judges in the panel); (c) discussion among judges in a panel about problematic items or about the cutscores in general; (d) performance data (item-level estimates of the probability that examinees will succeed on each item, usually in the form of \( p \)-values); and (e) intrajudge-consistency data, which can be presented as numerical information (resembling performance data feedback when presented as conditional \( p \)-values) or non-numerical, descriptive information (Loomis, 2000; Reckase, 2001; Reckase & Chen, 2012). In this study, I examine non-numerical intrajudge-consistency feedback and its effect on reducing error (intrajudge and interjudge inconsistency) in panelists’ judgments.

Error in panelists’ judgments can be examined at the individual judge level using measures of intrajudge inconsistency (Plake, Melican, & Mills, 1991; van der Linden, 1982) and at the panel level using a generalizability-theory framework (Brennan, 1995; Brennan & Lockwood, 1980; Kane, 1994), which provides data for evaluating the degree of interjudge inconsistency. Evidence of high intrajudge consistency (low intrajudge inconsistency) and low interjudge inconsistency in deciding where to place the cutscores contributes to the validity argument of the standard-setting process (Hambleton & Pitoniak, 2006; Hambleton et al., 2012; Kane, 1994).

**Measures of Intrajudge Consistency**

An assumption in Angoff standard setting is that judges have a consistent mental representation of both the construct and the operationalization of the construct (Hambleton & Pitoniak, 2006; Raymond & Reid, 2001; Reckase, 2006; van der Linden,
1982). That is, a judge should have a stable conceptualization of the BPE’s knowledge, skills, and ability (hereafter, ability), \( \theta_{BPE} \), and of the probability of success on each item given that BPE’s ability, \( p_i|\theta_{BPE} \). Measures of intrajudge consistency provide an estimate of the degree to which this assumption has been met (Camilli et al., 2001; Hambleton & Pitoniak, 2006; Kane, 1994, 2001).

Low intrajudge consistency may be due to lapses in attention, superficial examination of the items, fatigue, or unfamiliarity with the rating task (Chang, 1999; Plake et al., 1991; Smith & Smith, 1988). Inconsistency can also be due to judges’ lack of content knowledge on the item (Chang, Dziuban, Hynes, & Olson, 1996; Verheggen, Muijtjens, Van Os, & Schuwirth, 2008) or to unstable mental representations of the BPE (Plake et al., 1991). These unstable mental representations may be a result of low familiarity with the BPE’s context or of fluctuating conceptualizations of the BPE when the judge is asked to consider performance on different items.

**Between-round measures.** A seemingly straightforward but overly simplistic approach to evaluating intrajudge consistency would be to calculate correlations or deviations of a judge’s ratings on a set of items across two or more rounds (Berk, 1996; Plake & Impara, 1996). Because, however, the primary purpose of including multiple rounds of ratings is to have judges review feedback and improve judgments, a lack of change in ratings between rounds can occur if a judge is adhering to an agenda or is simply ignoring the feedback (Berk, 1996; Hambleton et al., 2012). Panelists who improve their judgments may, paradoxically, have low between-round correlations. Thus, this method is less than ideal for garnering internal validity evidence. A more suitable
approach to measuring intrajudge consistency involves examination of a judge’s ratings on items within each round.

**Within-round measures.** Hambleton and Pitoniak (2006) define intrajudge consistency as an indicator of “the degree to which a panelist is able to provide ratings that are consistent with the empirical item difficulties” (p. 458), where empirical item difficulties are obtained from a previous administration of the test with examinees from the same population for which the cutscore is intended to be used. Another perspective on this definition is presented by van der Linden (1982) and Plake, Melican, and Mills (1991), who describe low intrajudge consistency as occurring when a judge assigns a high probability of success on difficult items but a low probability of success on easy items.

**Conditional p-values.** Empirical item difficulties (such as $p$-values) are typically calculated based on the entire set of examinees attempting the item. A more appropriate measure, however, for evaluating degree of intrajudge consistency is the conditional item difficulty (Plake et al., 1991), which provides an estimate of an item’s difficulty at a particular level of examinee ability. In classical test theory, conditional $p$-values are calculated based on a group of examinees who scored at or within a band around the cutscore that operationalizes the BPE’s level of ability (e.g., Cizek & Fitzgerald, 1996; Goodwin, 1999). This calculation is not ideal, however, because there may be an inadequate number of examinees scoring within this band to get a stable estimate of BPE performance on each item (Hurtz & Jones, 2009).

**Item-response-theory conditional p-values.** In item response theory (IRT), conditional $p$-values can be calculated with greater precision than with classical test
theory methods because the point estimate, $\theta_{BPE}$, rather than a band, of the BPE’s ability is used in the calculation. The probability of success at the ability level is defined by the judge’s recommended cutscore on the latent-trait scale, which is represented in theta units (Hurtz & Jones, 2009; Sireci & Clauser, 2001; van der Linden, 1982). In other words, conditional $p$-values can be estimated based on each item’s item characteristic curve (ICC). With the one-parameter IRT (Rasch) model, the ICC is defined by

$$p_i|\theta_{BPE_j} = \frac{1}{1+\exp[\{b_i - \theta_{BPE_j}\}]}$$

where $p_i|\theta_{BPE_j}$ represents the probability of success on Item $i$ given $\theta_{BPE_j}$, which is Judge $j$’s estimate of where the BPE is on the latent-trait scale; and $b_i$ represents the IRT-estimated empirical difficulty parameter of Item $i$. In studies that employ this equation (e.g., Hurtz & Jones, 2009; van der Linden, 1982), $\theta_{BPE_j}$ represents an estimate of Judge $j$’s true intended cutscore (see Reckase, 1998, 2006), which is typically calculated based on the judge’s ratings across all items (e.g., Hurtz & Jones, 2009; van der Linden, 1982).

**Fit to the item characteristic curve.** Under IRT frameworks, intrajudge consistency at the item level has sometimes been referred to as the judge’s fit to the ICC (Hurtz & Jones, 2009; Hurtz, Jones, & Jones, 2008; Kane, 1987). That is, on the $Y$ axis of the ICC plot (where theta is on the $X$ axis and probability is on the $Y$ axis), an Angoff rating at approximately the same height as the empirical conditional probability (estimated from $\theta_{BPE_j}$) has high consistency; a large discrepancy between the Angoff rating and the conditional probability indicates low consistency (Figure 2.1).
Probability-scale deviations. Van der Linden’s (1982) seminal paper on intrajudge consistency employed this fit-to-the-ICC approach as the first step in estimating an intrajudge-consistency index. He designated the absolute value of the difference between the Angoff rating and the conditional probability the \textit{error of specification}, which is a measure of the lack of fit to the ICC. Van der Linden then transformed and inverted the error of specification to arrive at what he termed the \textit{consistency index}, a measure of intrajudge consistency. On this measure, higher numbers on the probability scale represent higher degrees of intrajudge consistency.
The transformation used in van der Linden (1982) involves the following steps:

1. After calculating the item-level error of specification, $ES_{ij}$, by Judge $j$ on Item $i$ as $ES_{ij} = |(\text{Angoff rating})_{ij} - p_i|\theta_{BPEj}|$, estimate the maximum possible error of specification, $MaxE_{ij}$, as $MaxE_{ij} = \max(p_i|\theta_{BPEj}, 1 - p_i|\theta_{BPEj})$.

2. Calculate the error of specification for the judge, $ES_j$, as the mean $ES_{ij}$ across items.

3. Calculate the maximum possible error of specification for the judge, $MaxE_j$, as the mean $MaxE_{ij}$ across items.

4. Calculate the consistency index of Judge $j$ as $C_j = \frac{MaxE_j - ES_j}{MaxE_j}$.

The consistency index places fit to the ICC on a scale that adjusts for the variation in empirical item difficulty (Chang, van der Linden, & Vos, 2004; van der Linden, 1982). In contrast, errors of specification do not adjust for empirical item difficulty; for example, an empirical conditional probability of .50 allows a maximum error of specification of only .50, while a value at .10 or .90 allows for a maximum error of .90 (if the judge provided a rating of 1.00 or 0.00 respectively). The adjustment made by the consistency-index calculation places judgments that are maximally different from the empirical item difficulty at the same value on the consistency-index scale (a consistency index of 0).

The consistency-index equation can also be used to estimate the consistency index of the judge at the item level (e.g., Chang et al., 2004): $C_{ij} = \frac{MaxE_{ij} - ES_{ij}}{MaxE_{ij}}$, where $C_{ij}$ represents the consistency index of Judge $j$ on Item $i$. Van der Linden (1982) suggested $C_{ij}$ can be utilized in providing judges with feedback about their degree of consistency on individual items. Chang, van der Linden, and Vos (2004) used this calculation to evaluate the validity of a method derived from Angoff and Nedelsky procedures.
Several studies (Chang, 1999; Chang et al., 2004; Ferdous & Plake, 2008; Friedman & Ho, 1990; Hurtz & Jones, 2009; van der Linden, 1982) have employed van der Linden’s (1982) equations to evaluate the standard setting procedures. Of these, four (Chang et al., 2004; Friedman & Ho, 1990; Hurtz & Jones, 2009; van der Linden, 1982) reported consistency indexes (summarized in Table 2.1).

Table 2.1

<table>
<thead>
<tr>
<th>Study</th>
<th>Test content</th>
<th>N items</th>
<th>N judges</th>
<th>Round</th>
<th>C</th>
<th>SD</th>
<th>Low</th>
<th>High</th>
</tr>
</thead>
<tbody>
<tr>
<td>Chang, van der Linden, &amp; Vos (2004)</td>
<td>German as a second language test</td>
<td>19</td>
<td>8</td>
<td>1</td>
<td>.748</td>
<td>.041</td>
<td>.68</td>
<td>.78</td>
</tr>
<tr>
<td>Friedman and Ho (1990)</td>
<td>Health certification exam</td>
<td>65</td>
<td>11</td>
<td>1</td>
<td>.745</td>
<td>.021</td>
<td>.72</td>
<td>.78</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>2</td>
<td>.775</td>
<td>.018</td>
<td>.74</td>
<td>.80</td>
</tr>
<tr>
<td>Hurtz and Jones (2009)</td>
<td>Simulated high-school proficiency data</td>
<td>75</td>
<td>13</td>
<td>1</td>
<td>.895</td>
<td>.003</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>2</td>
<td>.925</td>
<td>.002</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Van der Linden (1982)</td>
<td>High-school physics test</td>
<td>18</td>
<td>8</td>
<td>1</td>
<td>.769</td>
<td>.038</td>
<td>.71</td>
<td>.81</td>
</tr>
</tbody>
</table>

Note. C = mean consistency index across judges in the panel.

\(^a\)Chang, van der Linden, and Vos (2004) and van der Linden (1982) provided data for a single round; Friedman and Ho (1990) provided data for two rounds, one before and one after consensus feedback; Hurtz and Jones (2009) simulated the Round 2 data by manipulating the Round 1 simulated data to approximate the effect of various types of between-round feedback. \(^b\)Hurtz and Jones generated simulated data of 1000 iterations based on item parameters from a state high-school proficiency exam (from Hambleton, Swaminathan, & Rogers, 1991); they did not specify the content of the test.

Van der Linden (1982) demonstrated how the consistency index can be used to evaluate and compare standard setting procedures. He compared Angoff (N = 8 judges) and Nedelsky (N = 9 judges) panels’ intrajudge-consistency indexes on an 18-item high-school physics test and found, in the single round of rating, that the Angoff panel’s consistency index (M = .77, SD = .04) exceeded that of the Nedelsky panel (M = .68, SD = .04) by .11 probability units. Based on his demonstration, van der Linden suggested
using the consistency index when making decisions about judge selection, evaluating the effects of training and modifications (i.e., feedback), and in detecting items on which there was systematic judgment error. He also proposed the idea of providing judges with computer-mediated consistency-index feedback so that they can revise their item-level judgments.

Chang, van der Linden, and Vos (2004) used the consistency index to assess the functioning of a method they devised that combined components from Angoff and Nedelsky procedures. They reported consistency indexes ($M = .748, SD = .041$) of eight judges rating a 19-item German as a second language test in a single round.

Friedman and Ho (1990) investigated the effect of interjudge consensus feedback on intrajudge-consistency indexes on a 65-item health certification exam. From the first to the second (final) round, they found that all judges ($N = 11$) improved in their intrajudge-consistency indexes, with an increase from Round 1 ($M = .745, SD = .021$) to Round 2 ($M = .775, SD = .018$) of .030 probability units. They claimed that “the results showed that the techniques designed to improve consensus also improved consistency” (p. 10), but they did not include a control or comparison group, nor did they report an inferential statistical test.

Hurtz and Jones (2009) conducted a Monte Carlo simulation study to compare the stabilities of van der Linden’s (1982) consistency index and two other measures of fit to the ICC, one presented by Kane (1987) and a classical-test-theory approach by Goodwin (1999). They found van der Linden’s consistency index to be the most stable, even when simulated judges’ $\theta_{BPEj}$ estimates were set at relatively extreme locations on the theta
scale. Based on their findings, they recommended van der Linden’s equations for evaluating the accuracy of judges’ ratings in relation to items’ empirical item difficulties.

Hurtz and Jones (2009) simulated their second round of rating based on their speculated estimate of what could be the effect of various types of feedback on the Round 1 ratings. The increase in intrajudge consistency from Round 1 to Round 2 was the same as that reported in Friedman and Ho (1990), of about .030.

These studies (Chang et al., 2004; Friedman & Ho, 1990; Hurtz & Jones, 2009; van der Linden, 1982) support the use of van der Linden’s (1982) consistency index in evaluating standard-setting procedures. Across the three empirical studies [excluding Hurtz and Jones’ (2009) simulation] presented in Table 2.1, the mean Round 1 consistency index was .75 and the mean standard deviation was .03. The range of the observed consistency indexes of the 27 participants across the three studies was from .68 to .81, which provides a comparison for other Angoff studies using intrajudge-consistency indexes to evaluate the procedure.

**Theta-scale deviations.** Van der Linden (1982) also mentioned the possibility of measuring intrajudge inconsistency as the spread of a judge’s theta estimates across items, where a judge with high variability is estimated to have low intrajudge consistency. This approach has been mentioned in other publications (Cizek, 1996; Reckase, Bay, & ACT Inc., 1998; Taube, 1997) and entails first using the ICC equation to solve for the judge’s estimated BPE ability on Item $i$, $\theta_{BPEij}$, where the Angoff rating is used in place of the conditional probability (described in Plake & Cizek, 2012); that is,

$$\theta_{BPEij} = b_i - \ln \left[ \frac{1}{(\text{Angoff rating})_{ij}} - 1 \right].$$

Taube (1997) used this theta-scale deviation approach, calculating intrajudge inconsistency as the standard deviation of a judge’s $\theta_{BPEi}$. He did not report the intrajudge
inconsistency results, but used them to examine the relationship between intrajudge inconsistency and accuracy for 13 judges on several social-work licensure exams. He measured each judge’s accuracy as the correlation between empirical item difficulty estimates (IRT $b$-parameters) and the estimated item difficulty estimates (estimated $b$-parameters based on the ICC equation with Judge $j$’s Angoff ratings and $\theta_{BPEj}$). Taube found a strong negative correlation ($r \leq -0.80$, $p < .01$) between intrajudge inconsistency and accuracy, suggesting that “judges who held a more stable conception of minimal competence tended to estimate the performance of minimally competent examinees with greater accuracy” (p. 492). Taube suggested using this intrajudge consistency in making decisions about weighting individual judges’ ratings or in eliminating inconsistent judges.

Wyse (2009) also used a theta-scale deviation approach (though he did not use this terminology) in evaluating the consistency of 20 Angoff judges on the 2005 Grade 12 NAEP mathematics test. He analyzed absolute deviations on NAEP’s scoring scale, which was a linear transformation from the theta scale, on three rounds of rating. Judges participated in several between-round feedback activities, including a review of Reckase Charts, which provide intrajudge-consistency data to judges, after Round 1. Wyse reported a decrease in intrajudge consistency across the three rounds; for example, at the proficient-level cutscore, $M_{\text{round1}} = 23.40$, $M_{\text{round2}} = 14.82$, $M_{\text{round3}} = 12.19$. Because these estimates are on the NAEP scoring scale, their meaning is not directly communicable to researchers seeking to compare indexes of intrajudge consistency; nonetheless, it does show that judges participating in iterative rounds of rating improve in their consistency.

With a theta-scale deviation approach, a judge’s degree of inconsistency at the item level can be estimated as the difference between $\theta_{BPEij}$ and $\theta_{BPEj}$, similar to the procedure
Wyse (2009) employed, or as the standard deviation of the judge’s $\theta_{\text{BPEij}}$, as conducted by Taube (1997) and recommended by Cizek (1996). Communicating how much deviation is too much or not enough, however, appears to be restricted to individuals familiar with the scoring scale.

**Interpretability of intrajudge-consistency measures.** Measures of intrajudge consistency are important for evaluating the standard-setting procedure and establishing confidence in the cutscore recommended by the standard-setting panel (Hurtz & Jones, 2009). The probability-scale and theta-scale-deviation approaches both provide measures of intrajudge consistency and both have been used in standard-setting procedures. A merit of van der Linden’s (1982) consistency index is its interpretability, as it is in probability units. Whereas theta-deviations are on latent-trait scales that are unique to the test being examined, the consistency index allows for comparison across studies (van der Linden, 1982). The data in Table 2.1, while limited to only a few empirical studies, may provide comparisons for studies examining this measure. To agencies or other stakeholders seeking information on the degree of intrajudge consistency resulting from changes in the standard-setting procedure, the consistency index will likely serve as an informative statistic.

**Measures of Interjudge Inconsistency under a Generalizability-theory Framework**

Credibility of the panel’s recommended cutscore depends, in part, on the degree to which judges converge in their probability estimates (Kane, 1994). Under a generalizability-theory (G-theory) framework (Brennan, 2001; Cronbach et al., 1963; Shavelson & Webb, 1991) with an item-by-judge design, interjudge inconsistency can be composed of between-judge variability and item-judge interaction variability. Between-
judge variability comprises deviations among judges in their recommended cutscores on the test; that is, differences among judges after each judge’s mean Angoff ratings across all items has been calculated. Item-judge interaction has been labeled by some authors (e.g., Hurtz & Hertz, 1999) as intrajudge variability, but it should not be confused with intrajudge inconsistency. Item-judge interaction variability depends on the item-level ratings of other judges in the analysis (deviations between every judge’s item rating and the grand mean of all item ratings), whereas a judge’s intrajudge inconsistency (van der Linden, 1982) is independent of the ratings of other panelists.

One of the statistical assumptions in G theory (in which judges, or raters, are a random facet) is that raters in the study are a random selection from the domain of raters. In practice, random sampling of raters is seldom achieved in G-theory studies and researchers often make the case that the sample approximates a representation of the larger domain (Haertel, 2006). Random-sample recruiting in standard setting is likewise nearly impossible (Loomis, 2012), yet many Angoff studies (e.g., Brennan & Lockwood, 1980; Clauser, Harik, et al., 2009; Clauser, Swanson, & Harik, 2002; Fehrmann, Woehr, & Arthur, 1991; Fowell, Fewtrell, & McLaughlin, 2008; Hurtz & Hertz, 1999; Norcini, Lipner, Langdon, & Strecker, 1987; Verhoeven et al., 1999; Verhoeven, Verwijnen, Muijtjens, Scherpber, & van der Vleuten, 2002) employ a generalizability framework as it provides a means to examine variance components for evaluation or comparison purposes. Furthermore, the notion of exchangeability, discussed in Shavelson and Webb (1981), can be used to assume that a representative sample of a facet (such as judges) that is not necessarily randomly selected can provide reasonable variance-component information for considering the facet to be random.
**Variance components.** A generalizability study (G study) provides a means of estimating how much variance is explained by the components in a round of Angoff rating, including that of the items, judges, and item-by-judge interactions, which are confounded with unexplained variability. Several Angoff studies with an item-by-judge design (Brennan & Lockwood, 1980; Clauser, Harik, et al., 2009; Cross, Impara, Frary, & Jaeger, 1984; Hurtz & Hertz, 1999; Norcini et al., 1987; Verhoeven et al., 1999; Verhoeven et al., 2002) report variance components when evaluating or comparing features in the study. Among these studies, four (presented in Table 2.2) reported variance components of Angoff rounds before judges participated in any discussion or feedback activities. Three of the four studies reported in Table 2.2 were mostly in the field of medicine and are not optimally generalizable to education settings; nonetheless, they demonstrate that unexplained, item-by-judge, variability tends to make up a large portion of Angoff judges’ ratings.

Hurtz and Hertz (1999) examined eight large-scale occupational-licensure Angoff studies, in which judges participated in discussion of items, and found that on average, item variability made up 31%, judge variability made up 14%, and the interaction made up 56% of the total variance. More recent investigations (e.g., Clauser, Harik, et al., 2009; Clauser et al., 2002; Dillon & Walsh, 2000) have used G studies to examine the effect of feedback, such as performance-data feedback and discussion, on the change in variance components. I describe the effect of performance-data feedback on variance components in the item-level-feedback section below.
Table 2.2

Percent of Variance and Dependability Indexes in Pre-feedback Conditions in Item-by-judge Angoff Studies

<table>
<thead>
<tr>
<th>Study</th>
<th>Test content</th>
<th>N items</th>
<th>N judges</th>
<th>Percent of variance items judges</th>
<th>Percent of variance items-by-judges</th>
<th>Dependability index</th>
</tr>
</thead>
<tbody>
<tr>
<td>Brennan and Lockwood (1980)</td>
<td>Health</td>
<td>126</td>
<td>5</td>
<td>18</td>
<td>4</td>
<td>79</td>
</tr>
<tr>
<td>Clauser, Harik, et al. (2009)^a</td>
<td>Medical licensure</td>
<td>200</td>
<td>6</td>
<td>40</td>
<td>3</td>
<td>57</td>
</tr>
<tr>
<td>Cross et al. (1984)</td>
<td>Math teacher certification</td>
<td>195</td>
<td>6</td>
<td>40</td>
<td>3</td>
<td>57</td>
</tr>
<tr>
<td>Norcini et al. (1987)^b</td>
<td>Elementary teacher certification</td>
<td>60</td>
<td>5</td>
<td>16</td>
<td>15</td>
<td>69</td>
</tr>
<tr>
<td></td>
<td></td>
<td>75</td>
<td>5</td>
<td>12</td>
<td>52</td>
<td>36</td>
</tr>
<tr>
<td></td>
<td></td>
<td>90</td>
<td>6</td>
<td>42</td>
<td>10</td>
<td>47</td>
</tr>
</tbody>
</table>

Note. Estimates of the percent of variance of the components were calculated using variance components reported in the studies.

^aThe same six judges participated in the two Angoff procedures; the different sets of items measured the same content. ^bJudges in this study had been involved in the item development, which may have affected their degree of convergence in item ratings.

**The dependability index.** The dependability index is also used in evaluating Angoff procedures (Brennan, 2001; Raymond & Reid, 2001; Shavelson & Webb, 1991). It is calculated using the estimates of the variance components. Hurtz and Hertz (1999) used this measure to estimate the degree of judge consensus. They found, in the eight studies they examined, a range of dependability indexes from .66 to .87 (M = .76, SD = .08), which they labeled as “low” to “quite high” (p. 889). Dependability indexes of the data reported in Table 2.2 ranged from .41 to .82. These studies had six or fewer judges; estimates would likely increase with more judges. The dependability index provides an estimate of the degree of consistency among judges in their ratings of the items when we are not concerned with how much more difficult one item is than another.
Measures of the standard error of the cutscore. Variance components can be used to estimate the standard error of the cutscore if the procedure were to be replicated (Brennan, 1995; Raymond & Reid, 2001). According to the Standards (AERA, APA, & NCME, 1999), “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). Standard error estimates of the panel’s cutscore can be used (sometimes in conjunction with the test’s standard error of measurement) to identify a confidence interval around the cutscore (Cizek & Bunch, 2007; Fowell et al., 2008; Jaeger, 1991; Kane & Wilson, 1984). This information is crucial when standard-setting agencies determine the operating cutscore based on, along with policy concerns and stakes, judges’ recommended cutscore (Cizek & Bunch, 2007; Loomis, 2012).

Among several Angoff studies that have used G theory to estimate error variability (e.g., Clauser, Harik, et al., 2009; Clauser et al., 2002; Cross et al., 1984; Fehrmann et al., 1991; Fowell et al., 2008; Hurtz & Hertz, 1999; Verhoeven et al., 1999; Verhoeven et al., 2002), the most common equations for estimating the standard error of the cutscore have been those presented in Brennan and Lockwood (1980), who provided three measures that can be used in calculating the variance of the cutscore. The first measure, \( \sigma^2_X = \frac{\sigma^2_I}{n_i} + \frac{\sigma^2_J}{n_j} + \frac{\sigma^2_{IJ}}{n_in_j} \) (Brennan & Lockwood, 1980, Equation 8), provides an estimate of the variance of the cutscore under the assumption that judges have been sampled from the population of all possible judges and items have been sampled from the domain of all possible items. This calculation has been used by several studies (e.g., Cross et al., 1984; Fehrmann et al., 1991; Hurtz & Hertz, 1999). The second measure, \( \sigma^2_{X|I} = \frac{\sigma^2_I}{n_j} + \frac{\sigma^2_{IJ}}{n_in_j} \).
(Brennan & Lockwood, Equation 10), estimates error under the assumption that judges are a random facet but that items are a fixed facet. This estimate has also been used in Angoff standard-setting studies (e.g., Clauser, Harik, et al., 2009; Fowell et al., 2008; Verhoeven et al., 1999; Verhoeven et al., 2002). The third measure, $\hat{\sigma}_{X|J}^2 = \frac{\hat{\sigma}_j^2}{n_j} + \frac{\hat{\sigma}_{ij}^2}{n_i n_j}$ (Brennan & Lockwood, Equation 9), is less common, probably because it assumes items are random but judges are fixed; policy makers likely prefer to know the confidence limits around a cutscore if the procedure were replicated with similar, not the exact same, judges. The square root of each of these three measures provides the respective standard error of the cutscore.

_Treating random item components as fixed_. In estimating cutscore standard error, including items as a random component is appropriate when the items have been selected from a larger domain of possible items (Kane, 1994; Kane & Wilson, 1984). Item variability has implications for cutscore precision because it accounts for differences among test forms when they comprise different sets of items and differ in their overall difficulty (Kane, 1994). However, even when items on the standard-setting set are drawn from a larger sample, the item facet can be treated as fixed when (a) one wishes to estimate the standard-error of the cutscore with the specific items, or (b) it is assumed that the cutscore will be translated onto a latent trait scale to be used with other, equated, test forms (Clauser, Harik, et al., 2009; Clauser et al., 2002). The notion of exchangeability (Shavelson & Webb, 1981) can also provide reasonable grounds for examining a random facet as if it were fixed if the items were selected to represent the larger domain. Treating items as a random component would provide a conservative estimate of the standard-error of the cutscore; treating items as a fixed facet may prove
useful for evaluating the specific procedure and for situations in which the cutscore will be translated onto an IRT scale for equating to other forms of the test (Clauser, Harik, et al., 2009; Clauser et al., 2002).

**Estimating the minimum number of judges needed to meet a precision criterion.**

In a decision study (D study), researchers can examine the estimated effect that varying numbers of judges will have on the degree of precision in the cutscore. Hurtz and Hertz (1999), who used Brennan and Lockwood’s Equation 8 (treating items as random) with occupational license exams, conducted a D study to see the minimum number of judges needed to achieve a standard error of less than 2.5 raw-score points (about .015 probability points). They estimated that having 15 judges would have yielded their desirable criterion. Verhoeven, Verwijnen, Muijtjens, Scherpplier, and van der Vleuten (2002), who used Brennan and Lockwood’s Equation 10 (treating items as fixed) with a medical school exam, estimated the number of items and judges needed to achieve a standard error of .01 probability points. They found that 32 judges rating 200 items would have met their criterion. Estimates in both the Hurtz and Hertz and Verhoeven et al. studies were based on ratings made by judges who had participated in item-level discussion; thus, judges’ ratings were likely not statistically independent.

**Using Intrajudge and Interjudge Consistency Data to Evaluate Cutscore Credibility**

Standard-setting researchers (e.g., Camilli et al., 2001; Hambleton, 1978; Hambleton & Pitoniak, 2006; Popham, 1978; Scriven, 1978) accept that recommended cutscores will always be subjective, human judgments. What must be avoided are random, or capricious (Popham, 1978), judgments of the BPE’s ability. G theory provides interjudge-consistency information that is useful in examining how precise a cutscore is and how
many panelists would be needed to make the case that the subjective judgments are stable across judges. Intrajudge-consistency information indicates how well judges maintain a stable representation of the BPE across items, regardless of the variability among judges. Both of these measures, therefore, aid in evaluating the credibility of the procedure.

G studies provide a means of evaluating an Angoff procedure by informing analysts of the judges’ degree of convergence. Large discrepancies among judges in their cutscores, as indicated by large variability in the judge facet, might suggest capricious judgments but can also be attributed to judges’ systematic differences in their conceptualizations of the BPE. That is, they may differ in their subjective values while still being careful in making judgments. A high degree of item-judge variability is more serious, as it would suggest judges’ are idiosyncratically applying their BPE conceptualizations across items—at least as measured by the pattern of judgments made by all panel members. It is possible, however, that judges with different backgrounds and expertise attend to specific features within items and still make careful judgments; thus, both random variability (possibly due to capricious judgments) and systematic judge-specific variability (due to expertise and experience) can account for this variance component. Feedback to individual judges about how aligned their individual item judgments are with their overall judgments (item-level intrajudge-consistency data) would likely reduce this type of error.

Measures of intrajudge consistency aid in evaluating capriciousness by revealing irregularities relative to empirical item characteristic curves. Intrajudge-consistency data do not depend on judges’ subjective, value-based judgments. Instead, they tell us how well judges adhere to their BPE conceptualizations across items. Thus, a panel of judges
can potentially have a high degree of intrajudge consistency but a low degree of convergence in where to place the cutscore. Similarly, when a panel’s aggregate item-level estimates do not strongly correlate with the empirical item difficulty estimates, intrajudge consistency will provide different information from that supplied by a G-study’s item-judge variability estimate. Finally, intrajudge consistency provides no information about cutscore precision, but does assist in evaluating the credibility of the standard-setting procedure.

Additionally, as suggested by several authors (Plake et al., 1991; Reckase & Chen, 2012; van der Linden, 1982), information from estimates of item-level intrajudge consistency can be provided to judges, as item-level feedback between Angoff rounds, to inform them of where they may have made errors in the judgment process. This type of feedback permits judges to hold their own values but prompts them to carefully review items and conceptualize how well a BPE will do on each.

**Item-level Feedback**

A common approach to reducing judgment error has been to provide judges with empirical data in the form of item-difficulty reports, or $p$-values (Buckendahl, Smith, Impara, & Plake, 2002; Busch & Jaeger, 1990; Hambleton & Pitoniak, 2006; Plake & Impara, 2001; Plake et al., 1991; Reid, 1991). In the standard-setting literature, this type of feedback is identified as process feedback (Reckase, 2001) or performance-data feedback (Clauser, Harik, et al., 2009; Clauser, Mee, et al., 2009). The term *performance-data* suits the fact that $p$-values are estimates of examinees’ probability of success on each item based on empirical data collected from previous administrations of the test. From this perspective, the feedback informs judges when their ratings are not accurate
with respect to examinees’ performance (Cizek & Bunch, 2007). The term process feedback suits its purpose of informing judges when they may be misinterpreting the items and procedures (Hambleton & Pitoniak, 2006; Reckase, 2001; Reckase & Chen, 2012). Judges may have little experience making probability judgments and such data inform them of their judgment accuracy (Reckase & Chen, 2012). Judges are typically instructed to use this information to improve their judgment and to avoid blindly accepting the numerical data as their recommended item-level cutscores (Hambleton & Pitoniak, 2006).

Brandon (2004), in his review of the literature on Angoff standard setting, noted that studies providing traditional $p$-values as feedback (e.g., Plake & Impara, 2001) yielded higher accuracy estimates than studies not providing judges with these data (e.g., Chang, 1999). Other studies (e.g., Busch & Jaeger, 1990) reported increased correlations between Angoff ratings and $p$-values after judges were given $p$-values as feedback.

**Conditional $p$-value feedback.** Because traditional $p$-values are estimated based on the entire set of examinees, they may provide inaccurate information in that they over- or under-estimate the performance of the BPE group on the item (Cizek & Bunch, 2007). More correct information can be provided in conditional $p$-value feedback, which involves having judges examine reports of the $p$-values of examinees at or around one or more levels of examinee ability. Variations on this approach include provision of feedback at several, typically 10, possible bands of examinee ability or at multiple cutscore locations (or ability locations) on the test.

**Decile reports.** Several studies (Clauser, Harik, et al., 2009; Clauser, Mee, et al., 2009; Clauser et al., 2002; Dillon & Walsh, 2000; Margolis, 2011) use decile reports to
present judges with conditional $p$-value feedback. To generate these reports, examinees’ are rank-ordered and then categorized into 10 bands of proficiency; $p$-values for each proficiency band are then calculated. During the Angoff session, Judges are given, for each item, a table with the $p$-values at each of the bands of ability. Angoff studies investigating the effect of conditional-$p$-value feedback provided in decile reports (e.g., Clauser, Harik, et al., 2009; Clauser, Mee, et al., 2009; Dillon & Walsh, 2000) have found evidence suggesting this feedback results in an increase in (a) convergence among judges and (b) correlations between Angoff ratings and conditional $p$-values.

**The effect of conditional $p$-value feedback on variance components.** There is evidence suggesting that item-level process feedback reduces the variance among judges’ ratings. In Table 2.3, I list five studies in three articles (Clauser, Harik, et al., 2009; Clauser et al., 2002; Dillon & Walsh, 2000) that employ G theory to investigate the effect of conditional $p$-value feedback. The basic trend among these studies is that the percent of variance explained by judges, groups of judges, or error decreases after the feedback, suggesting improvements in convergence of judgments.

Conditional $p$-values provide more accurate feedback than traditional $p$-values calculated from the entire examinee set in that they inform judges about the performance of examinees at a particular level of ability. Even more accurate, however, is intrajudge-consistency feedback. The difference between conditional $p$-value feedback and intrajudge-consistency feedback is sometimes subtle; Reckase (2001) labels both as process feedback. In cases where the conditional $p$-values are estimated based on the individual judge’s estimate of the cutscore, $\theta_{BPEj}$, (rather than a priori determined deciles of ability) and judges are informed about how close their item ratings are to their intended $\theta_{BPEj}$, the data then serve as a kind of intrajudge-consistency feedback.
### Table 2.3

*Studies Reporting Variance Components before and after Conditional p-value Feedback*

<table>
<thead>
<tr>
<th>Study</th>
<th>Study Number</th>
<th>N judges</th>
<th>N items</th>
<th>Pre percent VC</th>
<th>Post percent VC</th>
<th>Change in percent VC</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>i  g  j  e</td>
<td>i  g  j  e</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Clauser, Harik, et al. (2009)*</td>
<td></td>
<td>6</td>
<td>34</td>
<td>70 0 30</td>
<td>80 0 20</td>
<td>+10 0 -10</td>
</tr>
<tr>
<td></td>
<td></td>
<td>6</td>
<td>43</td>
<td>68 1 31</td>
<td>84 1 15</td>
<td>+16 -83 -16</td>
</tr>
<tr>
<td>Clauser, Swanson, and Harik (2002)*</td>
<td></td>
<td>24</td>
<td>75</td>
<td>3 81 15 0</td>
<td>60 27 12 1</td>
<td>+57 -54 -3 +1</td>
</tr>
<tr>
<td></td>
<td></td>
<td>23</td>
<td>150</td>
<td>13 64 21 2</td>
<td>91 0 9 1</td>
<td>+78 -64 -12 -1</td>
</tr>
<tr>
<td>Dillon and Walsh (2000)</td>
<td></td>
<td>27</td>
<td>162</td>
<td>23 11 66</td>
<td>66 5 29</td>
<td>+43 -6 -37</td>
</tr>
</tbody>
</table>

*Note.* In studies in which group variance is reported, the judge variance is of judges nested within groups (j:g). VC = variance components, i = item facet, j = judge facet, g = group facet, e = error, which is ij or, with studies examining the group facet, i(xj:g).

*The judge variance had already been reduced in the pre condition due to a prior round of discussion on these items. This article’s first study, with 75 items, included not only conditional p-value feedback but also discussion, rater consensus, and consequence data. These variance components are from a D study based on the composite results of the multivariate analysis.*

**Intrajudge-consistency feedback.** Providing feedback on intrajudge consistency in Angoff methods has been repeatedly suggested (e.g., MacCann, 2009; Plake et al., 1991; Reckase & Chen, 2012; van der Linden, 1982). Research investigating whether this feedback has an effect on improving intrajudge and interjudge consistency appears yet to be reported in the mainstream standard-setting or measurement literature, however.

**Reckase Chart feedback.** One type of feedback that provides judges with item-level intrajudge-consistency feedback is often refereed to as Reckase Chart feedback [although not labeled as such in publications (e.g. Reckase & Chen, 2012) authored by Reckase himself]. With this feedback, judges are presented with the IRT-estimated conditional p-value data based on the judges’ rating behavior. Each judge receives a printout of this
chart where each column is an item and each row corresponds to an ability estimate (or a transformation of the ability estimate), with each step representing a scaled increase in ability. That judge’s Angoff rating on each item is highlighted to show its location and the degree to which that rating is not aligned with the rest of their ratings; perfect intrajudge consistency would result in all item ratings being located in the same row. In this way, Reckase Charts provide conditional \( p \)-value feedback at various locations on the theta scale and inform the judges where their ratings tend to occur. Judges can see, by looking at the average of their ratings on the total test score, where their initial recommended cutscore is and the conditional \( p \)-values at this level. Reckase Charts have been used as feedback in standard-setting procedures for the National Assessment of Educational Progress (NAEP). Examples of Reckase Charts are available in several publications (e.g., Loomis, 2000; Loomis & Bourque, 2001; Reckase, 2001; Reckase & Chen, 2012; Wyse, 2009; Yang, 2000).

**The effect of intrajudge-consistency feedback on variance components.** Reckase (2001) explains that “little solid research evidence is provided to show that feedback actually causes changes in ratings. The exception is work done by ACT to guide the design of the standard-setting process for NAEP” (p. 169). Reckase adds that providing intrajudge-consistency feedback to judges, using Reckase Charts, has had the effect of reducing variability among judges: “Process feedback [Reckase Charts] typically has the effect of reducing the standard deviation of cutscores set by panelists” (p. 169). A small between-judge standard deviation corresponds to a small variance component of the judge facet under a G-theory framework. Thus, this type of feedback may serve to reduce
the standard error of the cutscore by increasing convergence among judges. A
generalizability study would also shed light on the item-judge interaction effect as well.

Because intrajudge-consistency feedback, such as that provided in Reckase Charts, is
similar to conditional $p$-value feedback, and because conditional $p$-value feedback has
been shown to reduce the percent of variance attributed to the item-by-judge components
(as discussed in the conditional $p$-value feedback section above), it is reasonable to
assume that intrajudge-consistency feedback will have the same effect. Research
investigating this effect, however, appears to be lacking. Thus, there is a need for
empirical research investigating the effect of intrajudge-consistency feedback on
interjudge inconsistency, operationalized as the variance components of the judges and
item-by-judge interactions.

**Studies on the effect of intrajudge-consistency feedback.** I carried out a search on
EBSCOhost services, using Academic Search Premier and ERIC databases with searches
under the menus alltext, titles, and abstracts. With the intersection of keywords Angoff,
intra*, feedback (i.e., Angoff $\cap$ intra* $\cap$ feedback) and I found two studies (Friedman &
Ho, 1990; Plake et al., 1991). Neither of these empirically investigated intrajudge-
consistency feedback. Plake, Melican, and Mills (1991) provided a theoretical account of
the factors likely contributing to intrajudge consistency, and Friedman and Ho (1990)
examined the effect of judge-consensus feedback on judges’ intrajudge-consistency
indexes.

A search of the terms intra*, feedback, and standard* (i.e., intra* $\cap$ feedback $\cap$
standard*) turned up, among several articles unrelated to standard setting, two studies on
the topic (Loomis, 2000; Yang, 2000). The terms intra*, feedback, and cut* yielded these
same two publications. The Loomis and Yang studies both reported on the use of
Reckase Charts with the 1998 NAEP exams on civics and on writing. A search on
EBSCOhost services, using Academic Search Premier and ERIC databases with the terms
Reckase and feedback revealed no additional published studies investigating the effect of
intrajudge-consistency feedback.

Whereas Loomis (2000) and Yang (2000) described the positive self-reports of the
usefulness of Reckase Charts by judges, neither study was able to confirm that the
feedback had a direct effect on judges’ rating behavior. Yang explained that since the
NAEP standard-setting procedures did not permit experimental design, she was unable to
examine direct evidence of the effect of Reckase Charts on judges’ ratings. Reckase
Chart feedback was provided along with other feedback data and discussion, and
statistical control to detect its effect was not feasible. According to Reckase, “In those
field trials, it was not possible to consider feedback as a main effect with other aspects of
the process controlled. Therefore, there was confounding with other aspects of the studies”

After broadening the search using the terms Angoff and intra* (i.e., Angoff ∩ intra*)
and, in another search, Angoff and feedback (i.e., Angoff ∩ feedback), I was unable to
locate any additional empirical studies on the topic of intrajudge-consistency feedback.
Additionally, a search for articles citing Loomis (2000) and Yang (2000) in the Web of
Science’s Social Science Citation Index and in Google Scholar revealed no other studies
investigating the effect of intrajudge-consistency feedback.

I asked Mark Reckase, via email, if he knew of any studies conducting experimental
research on the effect of intrajudge-consistency feedback. His response was, “if by
experimental research you mean studies with random assignment and formal controls using other possible feedback mechanisms, the answer is that I know of no such studies” (M. Reckase, personal communication, May 30, 2011).

MacCann (2009) describes and recommends feedback that is nearly identical to Reckase Charts. He explains that “it has yet to be tested in practice whether such a process would result in greater consistency between judges than the usual Angoff method” (pp. 269–270). Thus, recent literature supports my conclusion that investigations into the effect of intrajudge-consistency feedback are scarce or nonexistent.

Aside from Loomis (2000) and Yang (2000), published research on the effect of intrajudge-consistency feedback is either nonexistent or not easily locatable in the usual channels. Experimental research into its effects appears to be lacking. In this study, I intend to fill this gap.

Given the calls for examination into the effect of item-level intrajudge-consistency feedback with Angoff judges (Eckes, 2009; MacCann, 2009; Plake et al., 1991; Reckase, 2001), but the apparent scarcity of documented experimental research in this area, there is a need for experimental studies examining its causal effect on judges’ ratings. Empirical data on whether intrajudge-consistency feedback improves judges’ rating behavior can contribute to researchers’ understanding of its value. Furthermore, practitioners considering this type of feedback can benefit from knowing whether it is worthwhile.

**Non-numerical intrajudge-consistency feedback.** Because of the similarity between conditional $p$-value feedback and intrajudge-consistency feedback that is presented numerically (e.g., as Reckase Charts) and because conditional $p$-value feedback has been shown to affect judges’ Angoff ratings (e.g., Clauser, Mee, et al., 2009), it is reasonable
to suppose that judges will use intrajudge-consistency feedback to change their ratings. However, a concern, raised by several authors (e.g., Clauser, Mee, et al.; Hoffman, Tashima, & Luck, 2010; Hurtz & Auerbach, 2003; Stahl & Becker, 2009; Taube, 1997), is that judges may be more likely to blindly adopt numerical feedback data, such as conditional \(p\)-values, as their Angoff ratings than to use the feedback to improve their judgment. This concern has been associated with the need to train judges in how to understand and use feedback data: “One possible outcome of poor understanding of the [feedback] data could be an unwelcome tendency to systematically change their ratings to be in line with data such as \(p\)-values, without considering the actual content of the item” (Hambleton & Pitoniak, 2006, pp. 455–456).

Proper training, though, may not be enough to ameliorate the tendency of judges to heavily rely on numerical data. Clauser, Mee, et al. (2009) provided evidence suggesting that Angoff judges blindly follow conditional \(p\)-value feedback regardless of their content expertise, thorough training, and having received explicit instructions to use the performance data to “…consider a second review of the item to see if there were content-related issues that they had not previously detected” (p. 393). The authors examined the rating behavior of 23 judges when they encountered feedback on 75 items, 36 of which had been manipulated so that the conditional \(p\)-value data feedback was incorrect by up to 1.5 standard deviations away from their empirical \(p\)-values. With these 36 items, judges’ post-feedback Angoff ratings were more highly correlated with the manipulated, incorrect, \(p\)-values \((r = .97)\) than with the empirical, correct, \(p\)-values \((r = -.05)\). Clauser, Mee, et al. characterized judges’ heavy reliance on the data as “somewhat mechanical” (p. 405), and suggested in their concluding remarks that alternative performance feedback
procedures need to be investigated. Research in descriptive, non-numerical, feedback may provide such an alternative.

Descriptive, non-numerical, item-level feedback has been used in Angoff standard setting procedures. According to Loomis (2000, 2012), ACT employed non-numerical intrajudge-consistency feedback in setting standards on the NAEP in the 1990s. With the 1995 NAEP Geography assessment, for example, feedback to each judge comprised two lists of five items on which the judge was most inconsistent: One list was made up of items the judge rated lower than expected and the other list comprised items the judge rated higher than expected. Each list was preceded by an explanation such as “Your overall rating average for the Proficient Level performance is 163.0. Given that, your ratings for these items are relatively low” (Loomis, 2000, Figure 12).

According to Loomis (2000), ACT stopped using descriptive intrajudge-consistency feedback because judges appeared to be ignoring or misunderstanding the data. Formal experimental investigation into the effect of this feedback, however, was apparently not carried out because the feedback was conflated with other between-round activities. In 1998, ACT began using Reckase Charts in lieu of descriptive intrajudge-consistency feedback because judges appeared to be using the Reckase Chart data to rate items in the subsequent round (Loomis, 2000). This is not surprising considering that Reckase Charts comprise numerical conditional probabilities and are therefore likely susceptible to the effect described in Clauser, Mee, et al. (2009).

How acceptable this blind acceptance of numerical data is, however, may depend on the stakeholders’ philosophical position. Reckase and Chen (2012) recommend that standard-setting planners consider where on the continuum between norm- and criterion-
referenced information the standard-setting process should be when designing the feedback procedures. That is, the extent to which judges are allowed to be guided by previous examinees’ performance should be considered when deciding the nature of the feedback. It follows that judges’ heavy reliance on numerical feedback may, therefore, be acceptable to stakeholders holding a more norm-referenced view. Investigators valuing a more criterion-referenced process or those using feedback to encourage judges to reconsider processes they go through while rating (Reckase, 2001), however, will likely perceive judges’ blind acceptance of numerical feedback as a threat to the validity of the standard-setting process.

Given that some stakeholders will perceive judges’ blind acceptance of numerical feedback as a threat to validity and that experimental research into non-numerical intrajudge-consistency feedback appears to be lacking, an investigation into non-numerical feedback is warranted. An experimental investigation into the effect of non-numerical item-level intrajudge-consistency feedback will likely contribute to the body of research in intrajudge consistency and provide empirical evidence that will aid practitioners when making decisions about which feedback activities to include in Angoff standard-setting procedures.

**Research Questions**

Considering that even though researchers have long recommended intrajudge-consistency data be given as feedback, but that empirical experimental research into whether it indeed has an effect on improving rating behavior is lacking, I ask the following research question:
1. Is there a difference in the changes in intrajudge consistency, operationalized as van der Linden’s (1982) intrajudge-consistency index, over the course of three rounds between a set of judges receiving non-numerical item-level intrajudge-consistency feedback and a set of judges asked to select and review items they believe they should reexamine?

This is the primary research question of my study. The null hypothesis is that there is no difference between the two groups in their changes in consistency indexes. My hypothesis, based on what we know about conditional p-value feedback and the self-report data in Loomis (2000) and Yang (2000), is that judges receiving the feedback will increase in intrajudge consistency significantly more than judges not receiving the feedback. However, because experimental evidence supporting this hypothesis is lacking, a second alternative hypothesis is possible: Judges asked to select their own items for reexamination might show significantly greater gains in consistency across the three rounds, which would suggest a detrimental effect of the feedback. Because both directions of alternative hypotheses are possible, I use two-tailed significance testing. I set the alpha level at .05. Greater improvements in the intrajudge-consistency indexes, across the three rounds, among the judges receiving feedback will support its use as a means to improve Angoff judge rating quality.

The next research question addresses interjudge inconsistency. This concerns the error variability due to the lack of convergence among judges in their Angoff ratings and in their resulting cutscores. I ask the following question:
2. To what degree does item-level intrajudge-consistency feedback appear to contribute to an improvement in interjudge consistency, as measured by changes in variance components, dependability indexes, and standard errors of the cutscore?

To answer this question, I conduct a G study on each of the three rounds of ratings for each of the two groups of judges—those receiving the feedback and those not receiving the feedback (where the judges not receiving the feedback are those asked to select and review items they believe they should reexamine). Similar to other Angoff studies using G theory (Clauser, Harik, et al., 2009; Clauser et al., 2002; Hurtz & Hertz, 1999), I examine (a) the percent of total variance explained by each of the variance components during each of the three rounds of ratings, (b) the dependability-index estimates, and (c) the standard errors of the cutscores. I treat items as the object of measurement and assume that judges are a random facet. To estimate the standard error of the cutscore, I use Brennan and Lockwood’s (1980) Equations 8 and 10; thus, I look at items as a random component and then as a fixed component.

The last research question is a follow up to Research Question 2, with a practical implication:

3. What is the minimum number of judges required to achieve an acceptable degree of precision?

This calls for a D study based on the G-study variance components estimated in addressing Research Question 2 and precision criteria used in previous standard-setting research. To answer this question, I estimate the precision in hypothetical scenarios (sampled from the same item domain and judge population as that from which my
observed data are assumed to be sampled) that differ in their numbers of judges.
Following Hurtz and Hertz (1999), I examine the standard error of the cutscore (when
items and judges are assumed to be random) and set the criterion to be a standard error of
less than 2.5 raw-score points. Following Raymond and Reid (2001), I also investigate
the dependability indexes and consider .80 as a plausible criterion. I investigate these in
the first and final rounds of the Angoff procedure to examine the effect of the intrajudge-
consistency feedback on the number of judges required to achieve the acceptable degree
of precision on these two criteria. Thus, the results will serve as an extension to Research
Question 2, with implications for practitioners and researchers planning standard-setting
procedures or conducting cost-benefit analyses with tests similar to the VST. The
Round 3 results will also contribute to the body of literature on the number of judges
required to achieve an acceptable degree of precision in Angoff standard-setting
procedures.

Because the panel’s recommended cutscore is the primary source of information that
standard-setting agencies ultimately use in determining where to place a cutscore, validity
of the standard-setting procedure is important. The credibility of a cutscore is
strengthened when the procedures are informed by empirical experimental research on
the effect of individual types of feedback such as intrajudge-consistency feedback.
Intrajudge-consistency indexes and interjudge inconsistency data are key sources of
evidence for establishing this validity and strengthening this credibility. The results from
my study will serve to inform researchers and practitioners of the value of item-level
intrajudge-consistency feedback.
CHAPTER 3

METHODS

Using a randomized experimental design, I developed standard-setting instruments and acted as the facilitator to carry out a simulated standard-setting procedure using a modified Angoff method. Participants were randomly assigned to either the treatment group, receiving intrajudge-consistency feedback, or the control group, participating in an equivalent activity with no feedback.

The standard-setting procedure differed from traditional Angoff methods in that the judges participated individually in the rating procedure and only reviewed one type of information between rounds. Having judges participate individually allowed for greater feasibility and flexibility in conducting an experiment with limited equipment, space, and time (as discussed in Cizek & Fitzgerald, 1996). Furthermore, whereas typical Angoff methods may include other data besides intrajudge-consistency feedback, such as discussion and interjudge consistency feedback (Cizek & Bunch, 2007; Hambleton & Pitoniak, 2006; Plake & Cizek, 2012), the method used in this study did not. The rationale for this was that the intrajudge-consistency feedback needed to be isolated from other sources of feedback that occur during standard-setting procedures in order to detect whether it had an effect, a limitation Reckase (2001) discussed in studies investigating feedback in standard setting.

Participants

Of the 36 adults that participated as standard-setting judges, 18 were in each of the two experimental (treatment and control) groups. All had at least an undergraduate degree and all had experience teaching at the secondary (Grades 6 through 12) or post-
secondary (college or university) level. Table 3.1 lists the percentage of participants with experience in these contexts. More than 70% (26) had taught middle- or high-school students, with about half (17) of the 36 participants having taught three or more years at this level. More than 80% (29) had college level teaching experience, with a third (12) of all participants having taught for more than three years. Slightly more than 40% (15) had experience teaching English or English-language arts, and 25% (9) could be classified as subject matter experts, given that they had three or more years of experience teaching this subject.

Table 3.1

<table>
<thead>
<tr>
<th>Teaching context</th>
<th>Number of participants&lt;sup&gt;a&lt;/sup&gt;</th>
<th>Years of teaching experience&lt;sup&gt;b&lt;/sup&gt;</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Less than 1</td>
</tr>
<tr>
<td>Middle- or high-school level</td>
<td>26 (72%)</td>
<td>4</td>
</tr>
<tr>
<td>University or college level</td>
<td>29 (81%)</td>
<td>5</td>
</tr>
<tr>
<td>English or English language arts</td>
<td>15 (42%)</td>
<td>2</td>
</tr>
</tbody>
</table>

<sup>a</sup>The percentages are out of the total number of participants, 36. Because many participants had experience in more than one teaching context, the percentages in this column sum to more than 100%. <sup>b</sup>This is the breakdown of the number of people with the years of experience in the context.

Table 3.2 shows the breakdown, within each of the two experimental groups, of the numbers and percentages of participants with relevant teaching experience. Among the 26 participants with experience teaching middle or high school students, 15 taught at the laboratory school where the vocabulary test that was used in the standard-setting procedure had been administered. That is, 42% of the participants had experience in the context in which the previous year’s vocabulary test data had come from.

The two experimental groups had approximately equal numbers of participants with experience teaching in each of the grade levels, as shown in Table 3.3. There was also
little difference between the two experimental groups in the numbers of participants with experience teaching in each college discipline (Table 3.4).

Table 3.2

*Number of Participants in Control and Treatment Groups with Relevant Teaching Experience*

<table>
<thead>
<tr>
<th>Teaching context</th>
<th>Within each experimental group&lt;sup&gt;a&lt;/sup&gt;</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Control (&lt;i&gt;n&lt;/i&gt; = 18)</td>
<td>Treatment (&lt;i&gt;n&lt;/i&gt; = 18)</td>
<td></td>
</tr>
<tr>
<td>Middle- or high-school level</td>
<td>14 (78%)</td>
<td>12 (67%)</td>
<td></td>
</tr>
<tr>
<td>University or college level</td>
<td>14 (78%)</td>
<td>15 (83%)</td>
<td></td>
</tr>
<tr>
<td>English or English language arts</td>
<td>5 (28%)</td>
<td>10 (56%)</td>
<td></td>
</tr>
<tr>
<td>Laboratory school</td>
<td>8 (44%)</td>
<td>7 (39%)</td>
<td></td>
</tr>
</tbody>
</table>

<sup>Note</sup>. Because many participants had experience in more than one teaching context, the percentages in each column sum to more than 100%.

<sup>a</sup>Percentages are out of the number within each experimental group, 18.

Table 3.3

*Number of Participants with Experience Teaching Each Secondary Level Grade*

<table>
<thead>
<tr>
<th>Grade</th>
<th>Within each experimental group&lt;sup&gt;a&lt;/sup&gt;</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Control (&lt;i&gt;n&lt;/i&gt; = 18)</td>
<td>Treatment (&lt;i&gt;n&lt;/i&gt; = 18)</td>
<td>Within entire sample&lt;sup&gt;b&lt;/sup&gt; (&lt;i&gt;N&lt;/i&gt; = 36)</td>
</tr>
<tr>
<td>Grade 6</td>
<td>8 (44%)</td>
<td>8 (44%)</td>
<td>16 (44%)</td>
</tr>
<tr>
<td>Grade 7</td>
<td>6 (33%)</td>
<td>8 (44%)</td>
<td>14 (39%)</td>
</tr>
<tr>
<td>Grade 8</td>
<td>6 (33%)</td>
<td>8 (44%)</td>
<td>14 (39%)</td>
</tr>
<tr>
<td>Grade 9</td>
<td>8 (44%)</td>
<td>10 (56%)</td>
<td>18 (50%)</td>
</tr>
<tr>
<td>Grade 10</td>
<td>10 (56%)</td>
<td>9 (50%)</td>
<td>19 (53%)</td>
</tr>
<tr>
<td>Grade 11</td>
<td>11 (61%)</td>
<td>9 (50%)</td>
<td>20 (56%)</td>
</tr>
<tr>
<td>Grade 12</td>
<td>11 (61%)</td>
<td>9 (50%)</td>
<td>20 (56%)</td>
</tr>
</tbody>
</table>

<sup>Note</sup>. Because many participants had experience in more than one grade, the percentages in each column sum to more than 100%.

<sup>a</sup>Percentages are out of the number within each experimental group, 18. <sup>b</sup>Percentages are out of the total number of participants, 36.
Table 3.4  
**Number of Participants with Experience Teaching Each College Discipline**

<table>
<thead>
<tr>
<th>Discipline</th>
<th>Within each experimental group</th>
<th>Within entire sample</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Control (n = 18)</td>
<td>Treatment (n = 18)</td>
</tr>
<tr>
<td>Social sciences</td>
<td>4 (22%)</td>
<td>6 (33%)</td>
</tr>
<tr>
<td>Sciences and math</td>
<td>5 (28%)</td>
<td>2 (11%)</td>
</tr>
<tr>
<td>College success</td>
<td>3 (17%)</td>
<td>4 (22%)</td>
</tr>
<tr>
<td>Arts and humanities</td>
<td>2 (11%)</td>
<td>2 (11%)</td>
</tr>
<tr>
<td>Technical skills</td>
<td>0 (0%)</td>
<td>1 (6%)</td>
</tr>
</tbody>
</table>

*Note.* The disciplines were coded based on teachers’ responses to the question that asked what subject areas teachers have taught to undergraduate students. Social sciences included education, psychology, assessment, and geography; sciences and math included biology, algebra preparation, and science teaching methods; college success included developmental education, composition, and test taking skills; arts and humanities included art, design, music, and civics; technical skills included database use and business writing.

All participants read and signed a consent form to participate. To ensure compliance with regulations and standards of ethical treatment of the participants, I followed the procedures specified by my institution’s Committee on Human Subjects for submitting my planned study for review by the committee. The committee approved my study.

**Materials**

**Test materials.** The test on which judges set the performance standard was a shortened version of the Vocabulary Size Test (VST; Beglar, 2010; Nation & Beglar, 2007). The full version of the VST comprises 140 items, but because the VST covers a very broad range of vocabulary knowledge, shortened versions can be administered when the examinees are expected to fall within a known range (Beglar, 2010). Beglar (2010) conducted a validity study on the full test and shortened, linked, versions when it was administered to both native and non-native speakers of English. He found strong validity
for its use to measure vocabulary size, and presented evidence for its unidimensionality and reliability (KR-20 = .96).

The item-response data used in my study came from a 2011 administration of a 103-item shortened version of the VST. The examinees were 310 students in intact classes in Grades 6 through 11. The test had been administered at a research laboratory school as part of a vocabulary growth study by its affiliated educational research organization. The laboratory school admits students using a stratified lottery system in order to approximate the demographics and prior academic achievement of the larger population of public school students in the state.

**Criteria for items selected for the standard-setting procedures.** I examined the Rasch modeled VST data, using Winsteps (Linacre, 2012), to identify whether to use the complete 103-item test or a subset of items. Table 3.5 lists the criteria and the evidence I examined.

**Criteria 1 and 2.** Rasch analysis of the 103-item test revealed that the first criterion was partially met. Whereas the reliability was high (KR-20 = .91), one item correlated negatively with the measure \( r = -.09 \). With Criterion 2, twenty-five items (24%) were estimated to have a standardized outfit index greater than 1.96. That is, nearly a quarter of the items were not measuring the test’s overall dimension as closely as the rest of the items on the test. Good item fit is important because the evaluation of judges’ consistencies requires that the test items already be internally consistent and unidimensional (van der Linden, 1982). Thus, I examined the remaining 78 items to see if they would serve as a viable subset for use in the Angoff procedures.
The 78-item set was reliable (KR-20 = .90), and none of the items correlated negatively with the modeled latent trait. Seven items (9%) exceeded the 1.96 standardized outfit criterion, which was three more than one would expect by chance (i.e., with 5% misfitting items); nonetheless, I proceeded with these 78 items with the caution in mind that feedback given to judges would reasonably, but not perfectly, align with the model intended to be used to score the students.

Table 3.5

<table>
<thead>
<tr>
<th>Criterion</th>
<th>Evidence examined</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Reliability above .80</td>
<td>KR-20 reliability; point-measure correlations(^a, b)</td>
</tr>
<tr>
<td>2. Reasonable fit of items to the Rasch model</td>
<td>Items with standardized outfit exceeding 1.96(^b)</td>
</tr>
<tr>
<td>3. Unidimensionality</td>
<td>Principal components analysis of residuals from the Rasch model(^b)</td>
</tr>
<tr>
<td>4. Absence of items with extreme scores</td>
<td>Items with (p)-value = 1.00 or (p)-value = 0.00(^b)</td>
</tr>
<tr>
<td>5. Normal distribution of item difficulty parameters</td>
<td>Quantile-quantile plot; significance test of non-normality(^c)</td>
</tr>
<tr>
<td>6. Approximately equal distribution of item difficulties of the practice and operational set of Angoff items</td>
<td>Descriptive statistics of item difficulty parameters in the practice and operational set(^c)</td>
</tr>
<tr>
<td>7. Number of items is large but not so large that judges become fatigued with the rating task</td>
<td>Pilot testing think-aloud data</td>
</tr>
</tbody>
</table>

\(^a\)Point-measure correlations reveal the strength and direction of the relationship between item functioning and the modeled latent trait; they are similar to point-biserial correlations (Linacre, 2009).  
\(^b\)Evidence was collected from Winsteps (Linacre, 2012) output. \(^c\)Evidence was collected using SAS PROC UNIVARIATE on the estimated item parameters from the Winsteps output.

Criteria 3 and 4. In examining the third criterion, I found that the second and third components in the Winsteps (Linacre, 2012) residual principal components analysis explained 2.1% and 1.9% of the un-modeled variance respectively, suggesting the 78-item test was reasonably unidimensional. With the fourth criterion, two items had
extreme scores; every examinee answered them correctly. Because these two items did not contribute to the model and their standard errors (1.83 logits for both) were very high, suggesting their difficulty estimates were unstable and not appropriate for feedback in this study, I removed them.

Criteria 5 and 6. With the 76-item set, probability plots and tests of non-normality suggested that the item difficulty parameters on the VST approximated a normal distribution: The quantile-quantile (Q-Q) plot approximated a straight line and the Shapiro-Wilk test provided no evidence to reject the null hypothesis that the data were normally distributed ($W = 0.98, p = .485$).

Scholars in the standard-setting literature (Clauser, Mee, et al., 2009; Plake & Cizek, 2012; Raymond & Reid, 2001) recommend that practice items used for training resemble the operational set on which judges will provide Angoff ratings. In my study, the items were reasonably similar to those I intended to use in the operational set primarily because they were drawn from the original 103-item test. The descriptive statistics of the difficulty estimates are presented in Table 3.6 and the normality of the data are presented in Table 3.7. Notable, perhaps, is that the distribution of the 10-item set ($SD = 2.54$) was similar to, but not equivalent to, the 76-item set ($SD = 1.84$). This difference likely existed because I had selected an extreme item (with an item-difficulty estimate of $p = 1.00$ probability of success due to all examinees getting it right) for the practice set to ensure that the participants had experience with a very easy item, as recommended by Taube (1997). The inclusion of this item resulted in leptokurtosis and negative skewness in the practice set, but not enough to suggest the data significantly diverged from normality (Shapiro-Wilk $W = 0.88, p = .141$). Also, interquartile ranges of the two sets
were similar, with that of the 76-item set being 2.56 logits and that of the 10-item practice set being 2.62 logits.

*Criterion 7.* I examined this criterion during the pilot-testing sessions, which I describe more fully in the data-collection section of this chapter. The think-aloud protocol data that I collected during the first two pilot-testing sessions suggested that participants were experiencing fatigue with the 76-item set during the three hours allotted to complete the task.

Because of this evidence of fatigue, I decided to select a subset of 50 items from this 76-item set. I determined that 50 items was adequate because other studies (e.g., Impara & Plake, 1998; van der Linden, 1982) examining Angoff procedures with educational tests also had 50 or fewer items. To ensure this subset of items represented the range of difficulty on the 76-item test, I sorted the 76 items by their difficulty parameters and removed every third item.

The distribution of item difficulty-parameters among these 50 items was very similar to that of the 76-item test, and moderately similar to those of the practice set, as indicated by the descriptive statistics in Table 3.6. I further pilot tested the 50-item set and concluded, based on the think-aloud data, that participants were not experiencing noticeable or self-reported fatigue with this 50-item set.
Table 3.6

*Descriptive Statistics of the Empirical Item Difficulty Estimates*

<table>
<thead>
<tr>
<th>Instrument</th>
<th>Items</th>
<th>Mean (SD)</th>
<th>Low</th>
<th>Q1</th>
<th>Median</th>
<th>Q3</th>
<th>High</th>
</tr>
</thead>
<tbody>
<tr>
<td>Early pilot-tested operational set</td>
<td>76</td>
<td>0.00 (1.77)</td>
<td>-3.62</td>
<td>-1.36</td>
<td>0.19</td>
<td>1.20</td>
<td>4.19</td>
</tr>
<tr>
<td>Operational Angoff set</td>
<td>50</td>
<td>0.10 (1.84)</td>
<td>-3.62</td>
<td>-1.36</td>
<td>0.26</td>
<td>1.28</td>
<td>4.19</td>
</tr>
<tr>
<td>Practice Angoff set</td>
<td>10</td>
<td>-0.49 (2.54)</td>
<td>-6.47</td>
<td>-1.47</td>
<td>0.93</td>
<td>1.15</td>
<td>2.55</td>
</tr>
</tbody>
</table>

*Note.* Units are in logits. The item parameters used in generating intrajudge-consistency feedback in the main study were taken from the Rasch model on the 76-item set. The practice Angoff set and the operational Angoff set were mutually exclusive. The median standard error of each of the three item sets was 0.13 logits.

Table 3.7

*Normality of the Empirical Item Difficulty Estimates*

<table>
<thead>
<tr>
<th>Instrument</th>
<th>Skewness</th>
<th>Kurtosis</th>
<th>Shapiro-Wilk test</th>
</tr>
</thead>
<tbody>
<tr>
<td>Early pilot-tested operational set</td>
<td>0.05</td>
<td>-0.35</td>
<td>0.98 .485</td>
</tr>
<tr>
<td>Operational Angoff set</td>
<td>0.02</td>
<td>-0.24</td>
<td>0.98 .655</td>
</tr>
<tr>
<td>Practice Angoff set</td>
<td>-1.48</td>
<td>2.97</td>
<td>0.88 .141</td>
</tr>
</tbody>
</table>

**Background questionnaire.** In order to collect data on the participants’ experience, I prepared an online background questionnaire in SurveyMonkey. To identify the relevant content, I utilized suggestions in standard-setting literature (Cizek & Bunch, 2007; Raymond & Reid, 2001). In developing the instrument’s instructions and items, I followed the guidelines in Dillman (2000) and Brown (2001).

In particular, the questions I developed were intended to identify the degree of knowledge and experience the participants had in (a) middle- and high-school teaching,
including whether such experience was in the local context, (b) college teaching, (c) language arts teaching experience and knowledge in language-learning theories, and (d) assessment, including whether they had experience in setting performance standards. I collected these data to characterize the participants, as recommended by Cizek and Bunch (2007).

Additionally, I intended to use the participants’ responses on subsets of items to create participant-background covariate measures that would provide estimates (scores) on three constructs: content expertise, assessment experience, and experience teaching in middle- or high-school contexts. These three constructs (content expertise, assessment expertise, and familiarity with the context of the examinees) are commonly cited in the literature (e.g., Brandon, 2004; Chang et al., 1996; Cizek & Bunch, 2007; Raymond & Reid, 2001; Reckase & Chen, 2012; Verheggen et al., 2008) as factors that affect judges’ ability to understand the procedure or to provide consistent ratings. Thus, I considered these to be important because they might be perceived as plausible alternative explanations of the treatment effect that I address in my first research question. In Appendix A, I provide the development outline of this instrument.

**Random-group assignment materials.** To carry out random experimental group assignment, I created a set of cards for each participant to blindly draw from at the start of each standard-setting session. Each card was coded, unbeknownst to the participant, as an experimental-group or treatment-group card using a unique randomly generated number. These numbers matched those in a table that only the standard-setting facilitator (the author) had access to and which assigned the participant to one of the two experimental groups. Because my plan was to conduct the experiment with 36
participants in a balanced design, I created 18 cards for the control condition and 18 cards for the treatment condition.

*Stratification materials.* One of the design criteria in my study was that at least 50% of the judges be participants with secondary teaching experience and that at least half of these be in each of the two experimental groups. To ensure these participants were represented in both experimental groups, I placed nine control cards and nine treatment cards in a bag labeled *rand-secondary*, from which participants with secondary experience drew, and I placed the remaining 18 cards in a second bag labeled *rand-other*. In other words, the two bags were identical in the number of treatment and control cards. The steps taken while using these materials to assign participants to the experimental groups are described in the procedures section below.

*Facilitator protocol checklist.* The standard-setting sessions conducted in this study occurred on several separate occasions in order to accommodate participants’ schedules and manage limited resources such as space and computers. To adhere as closely as possible to the same procedures across sessions, I developed a 21-step facilitator protocol checklist for checking off steps when carrying out the procedures. I developed this checklist using suggestions in Hambleton (2001) and refined the steps during the pilot-testing sessions, based on notes taken from participants’ think-aloud protocols. The facilitator protocol checklist is in Appendix B.

*Computers and software.* Windows machines were used in the sessions. All had Excel 2007 or 2010, with Visual Basic macros enabled. During the pilot-testing sessions, I inspected the standard-setting entry forms (described below) to ensure that all fields were visible on different machines with different monitors.
**Session observation log.** To document events that occurred in each session and identify unplanned occurrences that could have threatened procedural validity, I developed a facilitator session observation log. The log was designed for recording events at the participant level. It included fields for recording (a) the start and end time of each activity during the session, (b) the date and location of the session, (c) the computer the participant used, (d) the participant’s confidential code, (e) the random group-assignment number the participant selected, (f) the resulting experimental group to which the participant was assigned, and (g) fields for taking notes about what I observed or heard the participant say that might in any way be related to the standard-setting procedure’s validity. The observation log was developed and revised during pilot testing and is in Appendix C.

**Orientation sheets.** In order to train judges in the purpose and procedures in the standard-setting session, I developed an orientation sheet for each of the two experimental groups, drawing from Raymond and Reid’s (2001) guidelines in training judges. The control- and treatment-group orientation sheets were identical except in their descriptions of the between-round tasks, which differed between the two groups. Each orientation sheet explained the judge’s role and tasks, the purpose for setting the standard, and how the cutscore would be used. Also described was the definition of the barely proficient examinee, the test they would be reviewing, and an example item for rating. The orientation sheets were refined during the pilot-testing sessions using notes I had taken during participants’ think-aloud protocols. The experimental and treatment orientation sheets are in Appendix D.
Standard-setting entry forms. To implement the standard-setting procedures and record judges’ Angoff ratings and written comments, I developed four Excel files. Of the files, two were practice session entry forms, one for the treatment and the other for the control group, and two were operational session entry forms, again one for each of the two experimental groups. The practice session forms served as a process-training instrument to prepare the participants for the operational standard-setting procedure. The treatment and control group forms were identical except for the between-round activities, as well as the hidden formulas used for calculating and presenting the between-round information. Form titles did not reveal to participants to which experimental group the file belonged. The forms were pilot tested for accuracy in calculations and for verifying that the participants interacted with the interface in the intended manner. Representative excerpts of these instruments are presented in Appendix E.

Formatting and content. Each file comprised several pages, which were Excel sheets, with macro buttons for navigating to subsequent pages. Excel menus, sheet tabs, and sheets used in performing calculations (such as the intrajudge-consistency indexes) were hidden from the participants. All cells were locked for editing except for those into which the participants were asked to enter ratings and text. Spreadsheet cell backgrounds were systematically color-coded to make it clear to the participants which fields were for entering ratings, which provided information, and which were warnings or prompts. I programmed some cells to use conditional formatting for providing warnings and instructions to the participants if they entered invalid data or had not provided Angoff ratings for every item before moving to a subsequent page.
**Practice and operational form differences.** Because the practice session forms served as a process-training instrument, they differed in some ways from operational session forms. The practice session forms included (a) extensive instructions, which were on their own page and which had a button at the bottom that read “I have read these instructions” and directed the participants to the next page, (b) a page with instructions and practice in how to navigate the form’s interface, (c) a page for each of the two rounds of rating of the 10 practice-set items, and (d) between-round activity pages after the first round only. The operational session forms included brief instructions, three rounds with the 50 operational-set items, and between-round activities after the first and second rounds.

**Information for rating items.** Each round of rating within each form was on a single page and included the set of test items, a field for providing each item’s Angoff rating, and other information such as the item’s answer key and the rating the judge gave the item in the previous round. The instructions on the first round of rating on the operational session forms were, “Out of 100 high school seniors with barely enough vocabulary to successfully read college textbooks, how many will get the item correct? Enter a number for every item. Remember to look at the item carefully, including all possible answer choices, before you enter your rating. You can change your ratings for an item at any time while you are on this page.”

The answer key and the judge’s previous Angoff rating were conditionally formatted to be displayed only after the participant entered a rating for an item. Thus, for each item, the answer key and previous rating were not available until after the participant entered a rating, but after a rating had been entered, the participant was able to go back and change
the rating. The file did not, however, allow participants to revise ratings they provided in earlier rounds because they were unable to navigate to previous Excel pages.

Between each round were pages for the between-round information and activities, the content of which depended on whether the file was for the treatment or control group. Of the 50 items the judges rated, 12 items were selected for between-round activities because, as Reckase and Chen (2012) put it, “giving feedback on the judgments for every test item is impractical” (p. 153). I determined that 12 would be an appropriate number because it represented about 25% of the Angoff set, which I believed would not cause fatigue but which would provide a manageable set for the between-round activities to function as intended; furthermore, Friedman and Ho (1990) found that the 11 judges on their 65-item test made changes on approximately 12 items, suggesting that this may be a manageable number for judges to think about carefully. Pilot testing and think-alouds provided no data suggesting that 12 items were too many or too few.

**Treatment-group between-round pages.** The between-round pages for the treatment group provided intrajudge-consistency feedback to the participant. In the operational entry form, these pages included the following information: (a) the 12 items on which the participant gave the least consistent Angoff ratings in the preceding round, (b) the intrajudge-consistency feedback, which included information on the magnitude and direction of the inconsistency of the judge’s rating on the item in the previous round, (c) the rating the participant gave on each of these items in the preceding round, (d) the correct answer choice to each of these 12 items, and (e) a field for the participant to write comments. The instructions stated “You will see up to 12 items below. Your ratings on these items were not consistent with your ratings on the other items in Round 1. Carefully
reexamine each item as you review its feedback. Look at the word, the sentence, and the answer choices, and consider how the barely proficient students will answer the item.”

During the subsequent round of rating, the feedback that was provided and the comments the participant wrote were reproduced adjacent to the corresponding item.

Item-level intrajudge-consistency feedback. To provide feedback on the items the participant was least consistent in rating, three pieces of information were required: (a) the Angoff rating the participant gave on each of the 50 items during the round; (b) each item’s item difficulty estimate ($b_i$), which had already been calculated from the Rasch analysis on the 76-item VST; and (c) the test characteristic curve (TCC), which had also already been estimated based on the earlier Rasch analysis.

I programmed Excel to complete the following calculations after a treatment-group participant completed Round 1 and Round 2 (calculations were from van der Linden, 1982):

1. Calculate the participant’s (judge’s) recommended cutscore in raw units with the equation $\text{Cut}_{jr} = \sum_{i=1}^{n} p_{ijr}$, where $j =$ the judge, $i =$ the item, $r =$ the round, $n =$ 50 items, and $p_{ijr}$ represents Judge $j$’s Angoff rating divided by 100 (to transform into probability units) on Item $i$ during Round $r$. Thus, the sum of the probabilities can yield a recommended cutscore between 0 and 50.

2. Lookup the ability estimate that matched the judge’s recommended cutscore on the TCC; that is, find $\theta_j|\text{cut}_{jr}$ and rename this value Judge $j$’s recommended cutscore on the theta scale ($\theta_{jr}$).

3. Calculate the expected Angoff rating by Judge $j$ on Item $i$, given $\theta_{jr}$ and $b_i$ in the ICC equation, $E(p_{ijr}|\theta_{jr}) = \frac{1}{1+\exp[-(\theta_{jr}-b_i)]}$, where $E(p_{ijr}|\theta_{jr})$ represents the
expected Angoff rating in probability units (i.e., the conditional probability that
the barely proficient student at θjr will get Item i correct).

4. Subtract the observed Angoff rating from the expected Angoff rating,
\[ \text{Diff}_{ijr} = [p_{ijr} - E(p_{ijr}|\theta_{jr})]. \]

5. Calculate the error of specification (ES) as the absolute value of this difference,
\[ \text{ES}_{ijr} = |\text{Diff}_{ijr}|. \]

6. Calculate the maximum possible error of specification (MaxE) for the item given
the expected Angoff rating and the following logic: If \( E(p_{ijr}|\theta_{jr}) > .50 \), then
\[ \text{MaxE}_{ijr} = E(p_{ijr}|\theta_{jr}), \text{ else } \text{MaxE}_{ijr} = 1 - E(p_{ijr}|\theta_{jr}). \]

7. Calculate the consistency index (C) of Judge j on Item i in Round r as
\[ C_{ijr} = \frac{\text{MaxE}_{ijr} - \text{ES}_{ijr}}{\text{MaxE}_{ijr}}. \]

8. Sort the items by their consistency indexes, and select the 12 least consistent
items for feedback.

9. For each of the 12 least consistently rated items, create a description of the
magnitude of the inconsistency using the following logic: If \( C_{ijr} < .70 \), describe
magnitude as very; if \(.70 \leq C_{ijr} \leq .90 \), describe as somewhat; if \( C_{ijr} > .90 \), describe
as negligibly. I selected the .70 threshold as a rough estimate of excessive
inconsistency based on Round 1 consistency index data reported in empirical
studies (Chang et al., 2004; Friedman & Ho, 1990; van der Linden, 1982), where
no Angoff judge had a consistency index (across all items) lower than .68. I
selected .90 as a threshold because perfect consistency of \( C_{ijr} = 1.00 \) is highly
unlikely and the pilot testing revealed that \( C_{ijr} > .90 \) was almost never observed
among the 12 least consistent ratings; Furthermore, the .90 criterion is the same
as that used by Hurtz and Jones (2009), who considered .10 on the ES scale to be the maximum for a rating to be considered reasonably good fit to the ICC.

10. For each of the 12 inconsistent item ratings, provide the judge with feedback about the magnitude of inconsistency by displaying the following sentence, “Considering the way you rated the full set of items, your rating on this item was _____ inconsistent,” where the blank was replaced with very or somewhat, depending on the results of Step 9. If an item’s consistency index was greater than .90 (i.e., was negligibly inconsistent), no feedback was given for the item and the feedback field contained the text, “Skip this row. Your rating on the item that would have been shown in this row was consistent.” I included this latter conditional text because it would be unreasonable to provide corrective feedback on an item in which judges were already consistent.

11. Create a description of the direction of inconsistency using the following logic: If the observed Angoff rating was lower than expected (i.e., if $\text{Diff}_{ijr} < 0$), describe as more; if the observed Angoff rating was higher than expected (i.e., if $\text{Diff}_{ijr} > 0$), describe as fewer.

12. For each of the 12 inconsistent items, provide the judge with feedback about the direction of inconsistency by displaying the following sentence, “Considering the way you rated the full set of items, it is likely that _____ of the barely proficient students would have selected choice _____.” In place of the first blank, there were four possibilities: “MANY FEWER”, “FEWER”, “MORE”, and “MANY MORE”. The qualifier “MANY” was included using the same criterion to include “very” in the magnitude feedback. If the Angoff rating was lower than
expected, “MORE” appeared, and if the Angoff rating was higher than expected “FEWER” appeared. The letter of the correct answer choice to the item replaced the second blank.

13. For each of the 12 inconsistent items, provide instructions that direct the participant to reexamine the item and write comments in an adjacent textbox. Here is the written prompt: “Carefully reexamine each item as you review its feedback. Look at the word, the sentence, and the answer choices, and consider how the barely proficient students will answer the item. Writing comments can help you organize your thinking about the item after you have reviewed the feedback. If you cannot think of anything to write for an item, you can leave that item’s comment box blank.” This follows suggestions in van der Linden (1982) and Pitoniak (2003) that judges be given a chance to reexamine their thinking about item functioning based on the feedback.

14. Calculate the judge’s observed intrajudge consistency for the round (C_{jr}). This value was not provided as feedback to the judges, but was to be used as the outcome variable for the primary research question in the study. This procedure used the same equations as those reported in van der Linden (1982). C_{jr} was equivalent to calculating the mean C_{ijr} across all items for Judge j during Round r.

**Control-group between-round pages.** The between-round pages for the control group provided an alternative condition to that which the treatment group experienced. Whereas the treatment group had the 12 items automatically selected for them, the control group participants were asked to select 12 items on their own. Thus, after completing their ratings, the next page was an item-selection page. The instructions asked
participants to “select exactly 12 items which you will more closely analyze in the next section. These might be items which you are unsure about or which you have difficulty imagining the way a barely proficient high-school senior will respond.” All 50 items, along with the answer key, the Angoff rating the participant gave in the previous round, and a field for selecting the item as one of the 12 for closer inspection were reprinted on the page. The number of items the participant had selected was displayed at the top and a warning was displayed if more than 12 items were selected.

The second between-round page displayed the 12 items the participant selected and asked the participant to reexamine the items: The instructions were, “You selected these 12 items. Reexamine each item and write comments about it. Look at the word, the sentence, and the answer choices, and consider how the barely proficient students will answer the item. If you do not have anything to comment on about an item, you can leave that item’s box blank and move to the next item.” The answer key, the rating the participant gave the item in the previous round, and a field for the participants to enter comments were presented. During the subsequent round of rating, the comments the participant wrote were reproduced adjacent to the corresponding item.

**Post-session questionnaires.** I developed post-practice-session and post-operational-session questionnaires for each of the two experimental groups. These questionnaires, which were on paper, served as a verification that the participants were paying adequate attention to key components of the standard-setting sessions, which was important after the practice session, as it provided a means to identify key parts of the practice session, if any, that needed reviewing. The questionnaires asked participants the extent to which, on a scale from 0 to 10, they paid attention to each of the following: (a) the instructions,
(b) the vocabulary word in the stem of the item, (c) the item’s sentence, which provided
the syntactic context of the vocabulary word, (d) the distractors, (e) the correct answer
choice, (f) the feedback (treatment) or the comments the participant had written (control),
and (g) the previous round’s rating. These questionnaires were for immediate verification
during the standard-setting procedure rather than for analysis. The post-practice session
questionnaire is in Appendix F.

Data Collection

Pilot testing. Before conducting the main study, I pilot tested the instruments and
procedures with seven participants. The purpose of this was to identify and address
problems with logistics, instructions, and prompts. Logistics included the recruiting
procedures, computing hardware, Excel displays, online survey functioning, timing of
tasks in the standard-setting procedure, and facility use. Investigating instructions and
prompts included making sure the participants correctly interpreted the textual and
graphic prompts, followed the instructions, and were not fatigued by length of these
instructions and prompts.

Key features of the instruments that I investigated and the criteria I used to determine
their adequacy are listed in Table 3.8. More detailed descriptions of the instrument edits I
conducted during pilot testing are in the instrument and procedure development outlines,
in Appendix A.
Table 3.8
Features of the Instruments Investigated During Pilot Testing

<table>
<thead>
<tr>
<th>Instrument</th>
<th>Instructions</th>
<th>Prompts</th>
<th>Formatting</th>
<th>Formulas</th>
<th>Time</th>
</tr>
</thead>
<tbody>
<tr>
<td>Recruitment materials</td>
<td></td>
<td>x</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Background questionnaire</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Orientation sheets</td>
<td>x</td>
<td></td>
<td>x</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Facilitator checklist</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Session observation log</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Practice standard-setting entry form</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
</tr>
<tr>
<td>Operational standard-setting form</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
</tr>
</tbody>
</table>

*aThese instruments were not viewed by the participants, as they were for use by the facilitator.

To determine whether the instruments that the participants interacted with were functioning well and to identify where edits needed to be made, I conducted verbal (think-aloud) protocol procedures following the guidelines in Ericsson and Simon (1993). Instructions given to the participants to train them in the verbal protocol are in Appendix G. While the participants were verbalizing their thoughts, I recorded, using the observation log, comments and issues that arose. When verbalizations were made that I did not understand, I asked follow-up questions after the participant completed the procedures. Additionally, I examined the participants’ written responses on the standard-setting forms to ensure they responded to the prompts in the intended manner. To ensure the sessions did not extend over the three-hour limit I had set, I recorded the length of time the participants took in completing each portion of the experiment.
After each pilot-session, I reviewed my notes, paying to attention to the features and to any threats to validity that I had not considered beforehand. After the first two pilot sessions, I revised the instruments. For example, because the second participant was experiencing fatigue with the piloted 76-item set of Angoff items (and exceeded the three-hour time limit), I reduced the number of items to 50 by sorting the 76 items by their difficulty parameters and removing every third item. (No issues arose with time and fatigue in subsequent pilot sessions.) After each of the remaining five pilot sessions, I conducted minor revisions, correcting copyedit errors and formatting.

**Participant recruitment.** In literature providing guidelines for selecting participants to serve as judges on standard-setting panels (AERA, APA, & NCME, 1999; Cizek & Bunch, 2007; Hambleton et al., 2012; Raymond & Reid, 2001), authors advise standard-setting agencies to determine the size and representativeness of the desired sample before recruiting participants. My target was to recruit 36 participants and that at least 18 (50%) of these participants have experience teaching middle- or high-school students, at least 9 (25%) have experience teaching at the university or college level, and that at least 9 (25%) have experience teaching English or English language arts. That is, I set out to recruit participants who had experience with (a) the student population for which the standards would be set, (b) the context that the performance standard generalized to, which was success in college reading, or (c) the subject area of the performance standard. I expected that some participants would have experience in more than one of these three categories.

**Sample size.** My target to recruit 36 participants was so that I could assign 18 participants to each experimental group. I decided on the number 18 based on
recommendations in the literature and previous Angoff-method studies, as well as feasibility in recruiting, paying, and arranging times and places to meet participants. Table 3.9 lists recommendations in the literature for the number of judges to include.

With this information, I concluded that having 18 participants per experimental group reasonably met the recommendations and approximated the typical size reported in previous studies.

Table 3.9

*Recommendations on the Number of Judges on an Angoff Standard-setting Panel*

<table>
<thead>
<tr>
<th>Publication</th>
<th>Method</th>
<th>Number of judges</th>
</tr>
</thead>
<tbody>
<tr>
<td>Jaeger (1991)</td>
<td>Standard error of the mean of judges’ ratings compared to the test’s standard error of measurement</td>
<td>13</td>
</tr>
<tr>
<td>Hurtz and Hertz (1999)</td>
<td>Generalizability-theory (D-study) analyses</td>
<td>10 to 15</td>
</tr>
<tr>
<td>Raymond and Reid (2001)</td>
<td>Review of studies</td>
<td>10 to 20</td>
</tr>
<tr>
<td>Brandon (2004)</td>
<td>Review of studies</td>
<td>10 to 20</td>
</tr>
<tr>
<td>Hambleton and Pitoniak (2006)</td>
<td>Review of studies</td>
<td>15 to 30</td>
</tr>
</tbody>
</table>

**Recruitment emails.** To recruit participants, I sent email invitations through the listserv mailing lists of several local groups and to individuals. I selected these groups and individuals because of their accessibility and because there was a good chance I would successfully meet my target of recruiting a representative sample of stakeholders with experience in the two instructional contexts and in the subject matter. All individuals and groups included people residing on the island of Oahu.

**Group mailing lists.** The groups receiving email invitations included five university departments in the field of education, a group of secondary-school teachers who had attended a writing and vocabulary professional-development series, teachers at the
laboratory school where the vocabulary test had been administered, and a learning assistance center for undergraduates seeking tutoring services. The total number of people receiving the group email invitations through the mailing lists is unknown.

*Individual recruitment.* The individuals receiving email invitations were faculty at local community colleges and the public university, as well as individual professionals with whom I was familiar and who I believed would be qualified to serve as judges because of their experience and content knowledge. Additionally, I asked the participants to recommend any of their colleagues who they believed would be qualified because of their teaching experience or knowledge of language learning; this process yielded three recruited participants.

*Minimum and desirable qualifications.* In the recruitment email message, I specified that the minimum qualifications were that participants be over age 18, that they have an undergraduate degree, and that they consider themselves to be native speakers of English. This was to meet institutional review board requirements and to ensure that participants were sufficiently knowledgeable in the content domain of the test. The email also specified that I sought participants with experience teaching at the secondary or post-secondary level and that participants would receive $20 in exchange for their three hours of participation.

*Time of recruitment.* I sent out email invitations in several waves between the months of February and June, 2012. The first wave was sent to individuals and to the group of teachers at the local laboratory school where the VST had been administered. I recruited from this group first to ensure that I had exhausted sampling from the most
representative group—the teachers who had knowledge of the students for whom the standard would be set.

In later waves, I sent invitations through the university departments, the vocabulary professional-development list, the university’s Learning Assistance Center tutor list, and to individuals teaching developmental education and English courses at several of the local community colleges. I continued sending invitations until I was able to recruit 36 participants, a goal that I achieved in June, 2012.

Because sampling from the universe of all standard-setting stakeholders is typically expensive and challenging even for well-funded organizations (see Loomis, 2012), I did not attempt to randomly sample from the population of all possible local stakeholders. Thus, as with most standard-setting studies, generalization of my results to a larger domain must be taken with caution.

Nonetheless, apart from the continuous recruitment procedures, this study’s selection and recruitment procedures corresponded with Raymond and Reid’s (2001) recommendations for panelist recruitment. Furthermore, other standard-setting studies (e.g., Chang, 1999) have conducted convenience sampling with somewhat more expedient procedures, such as having students in an intact graduate level class serve as a panel of judges.

Background questionnaire data collection. After recruiting the participants, I sent follow-up correspondence emails for the purpose of (a) providing them with the consent form to read beforehand, (b) arranging a meeting time and place for the standard-setting session, and (c) providing a link to the online background questionnaire. The participants
were asked to complete the questionnaire before arriving at the standard-setting session. All participants completed it before the meetings.

The median amount of time the participants took to complete the questionnaire was six minutes. Responses to Question 2, which asked whether the person had experience teaching middle- or high-school students, were used to categorize the participants into one of the two strata used in the stratified group assignment procedure, which is described next.

**Random group assignment.** In this section, I first describe the plan I designed for assigning participants to one of the two experimental conditions. I then describe what I did and what the results of the random group assignment were.

**Group assignment plan.** In planning the random group assignment procedure, I consulted Shadish, Cook, and Campbell (2002) and considered three design criteria. First, I intended to achieve a balanced experimental design, with an equal number of participants in the control and treatment groups. This required that I use a restricted random assignment procedure, which I describe below. Second, I wanted to ensure that participants with secondary-teaching experience were minimally represented in both experimental groups; thus, I needed to include a stratification procedure. Third, I needed to account for the trickling in of participants because my recruitment was intended to be ongoing; therefore, I needed to plan procedures in the event that the probability of being in one of the two experimental groups became excessively lopsided as participants entered the study.

**Restricted random assignment plan.** Whereas simple random assignment procedures do not guarantee equal sample sizes per condition, a restricted random assignment
procedure does so by allotting an equal number of slots for participants in each condition. To ensure I would achieve a balanced design, I prepared the set of 36 random-assignment cards, which were evenly split to assign 18 participants to each experimental condition. The cards were to be drawn, without replacement, by participants at the start of each standard-setting session. Thus, the probability of selecting a control-group or treatment-group card depended on how many available cards remained in each category after previous sessions’ participants selected their cards.

**Stratification plan.** The primary stakeholder group in this study was secondary school teachers. My criterion was that at least half of the 36 participants be from this population and that at least nine (25%) be in the control and nine (25%) be in the treatment group.

To represent this population in both experimental conditions, I planned to assign the first 18 participants with this experience into a stratum labeled as *secondary-experience*. The remaining participants would be assigned to a stratum labeled as *other*, regardless of whether they had such experience. That is, participants who had secondary-school teaching experience but who had entered the study after the first 18 *secondary-experience* slots were filled would also be assigned to the *other* stratum. The strata’s bags, which I labeled as the *rand-secondary bag* and the *rand-other bag*, each initially contained nine control and nine treatment cards.

**Trickle-in plan.** Because my procedure entailed assigning participants as they entered the study, there was a risk that one of the two experimental-group card sets would be exhausted before the other. That is, because I had a restricted random assignment plan and planned to add the participants as they trickled in, there was a chance that the
probability of being in one of the two groups (treatment or control) would be 1.00 before I completed the recruitment. Such a scenario would threaten validity if the later sessions’ participants differed in some systematic way from earlier sessions’ participants (Shadish et al., 2002). To reduce the threat of systematic group assignment, I planned to include multiple participants per session as often as I could because these participants would be instructed to draw cards simultaneously, which would give each participant in the session the same chance of being in each of the two experimental groups.

Exhaustion of one of the experimental group assignments was possible in two scenarios: when no more cards for an experimental group are available overall, regardless of stratum, and when no more cards for an experimental group are available within one of the two stratum bags. An example of the first scenario would be if all 18 control-group cards, regardless of which stratum bag they came from, had been selected by incoming participants. This would force all subsequently incoming participants into the treatment group unless a plan was in place to generate more cards, which would adjust the probability of being in one of the two groups back to a less extreme level.

An example of the second scenario would be if all nine control-group cards had been selected from a stratum bag. This would force subsequent participants within this stratum into the treatment group unless a plan was in place to either generate more cards or combine all remaining cards into a single bag. Generating more cards would, of course, translate into higher costs in time and money. Therefore, with this second scenario, if one experimental group became exhausted before another within a stratum, my plan was to combine all remaining cards into the same stratum bag and to have incoming participants select from this single bag regardless of whether they had experience in secondary
teaching contexts. This could complicate the stratification plan but would be a reasonable tradeoff.

In the event that one experimental group was exhausted before the other (the first scenario described above), I planned to recruit up to four more participants and to generate a new set of cards for the remaining participants, with an equal number of control and treatment cards. This would complicate the plan to achieve a balanced design but would be a reasonable tradeoff.

**Group assignment procedures.** There were 23 standard-setting sessions in all, in which 10 (43%) were attended by only one participant and 13 (57%) were attended by two or three participants. Those attending the same session drew cards at the same time. All participants were asked to select from one of the two stratum bags at the very start of the standard-setting session. Each card had a random number on its face, which I then looked up on a prepared table that indicated to which experimental group the card belonged. I then recorded the participant’s experimental group on the session observation log and discarded the drawn card. The remaining standard-setting materials that each participant received depended on which group to which the participant was assigned.

Before holding each standard-setting session, I determined which stratum bag each participant should select from by viewing each participant’s response on the online background questionnaire item (Question 2) that asked whether he or she had experience teaching middle- or high-school students. Those with such experience were assigned to the *secondary-experience* stratum. Those without this experience or those participating after the first 18 *secondary-experience* slots were filled were assigned to the *other*
stratum. I recorded each participant’s stratum on the session observation log, which I then referenced during the standard-setting session.

**Group assignment results.** My stratification plan was to ensure that both experimental groups were adequately represented by participants with experience teaching middle or high-school students. My criterion to have at least 50% representation in each group was met: As Table 3.2 above shows, 78% of the control group (14 participants) and 67% of the treatment group (12 participants) were represented by participants with middle- or high-school teaching experience.

Among the potential problems associated with the trickling in of participants, only one instance occurred in which I was forced to adjust the probability. During Session 14, in which there was only one card left in the secondary-experience stratum, the last participant to select from this bag would have automatically been assigned to the treatment group had I not made the adjustment. Instead of forcing this participant to be in the treatment group, I placed this remaining card into the rand-other bag and had the participant draw from this bag, which then contained eight control cards (.47 probability) and nine treatment cards (.53 probability).

The probability of being assigned to each of the two groups was approximately equal throughout the study. Tables H.1 and H.2 in Appendix H provide the probabilities of being assigned to each group during each of the 23 sessions. For 19 (83%) of the 23 sessions, the probability of being in the treatment or control group was between .43 and .57. The most unbalanced probability occurred during Sessions 9 and 22, in which there was a .33 probability of being in the control group for each of these sessions.
**Standard-setting sessions.** I acted as a standard-setting facilitator and conducted the 23 standard-setting sessions between the months of February and June, 2012. Sessions took place in conference rooms, vacant classrooms, or in offices, depending on site availability and convenience for the participants. Each session occurred during a single meeting and was attended by either one, two, or three participants in each session, depending on the participants’ schedules. The time taken by the participants to complete the sessions was between 1 hr 20 min and 3 hr 00 min, $M = 2$ hr 03 min, $SD = 26$ min. For each of the two experimental groups, this length of time was similar: For the control group, the mean time taken was 2 hr 09 min ($SD = 24$ min); for the treatment group, the mean was 1 hr 57 min ($SD = 28$ min). The participants were asked to take a break after the practice session, usually between 5 and 15 minutes, and were told that they could step out at any time if they needed to take other breaks. No session was ended early because of time constraints. I provided beverages and snacks.

To maintain standardization across sessions, I following the scripted procedures outlined in the standard-setting facilitator protocol checklist provided in Appendix B. The following is an abbreviated version of the steps in the sessions:

1. After receiving a response from potential participants, I arranged a three-hour block of time to meet at a conference room, a vacant classroom, or other quiet location for the standard-setting session.

2. I emailed the participants the consent form and the link to the background questionnaire.

3. At the start of the standard-setting session, I assigned each participant to the treatment or control group depending on which card the participant selected.
4. In order to train the participants, I asked them to read the orientation sheet, which presented an outline of the session, explained the purpose and procedures, and described the barely proficient student.

5. After the participants completed the orientation sheet, I asked them to restate their role and purpose for setting the standard, to describe the barely proficient student, to describe the items on the test, and to ask any questions they may have about the procedures. This was to verify that participants were adequately trained.

6. I guided the participants to complete the practice session and encouraged them to ask questions at any time.

7. I administered the post-practice session questionnaire to verify that the participants understood the rating process, and I addressed questions.

8. I asked the participants to complete the operational session, which included three rounds and experimental group activities prior to Rounds 2 and 3.

9. I administered the post-actual session questionnaire.

10. I thanked and paid the participants.

11. Finally, I saved the completed standard-setting forms for later analysis.

The treatment and control group participants participated in the same activities except for those tasks occurring between rounds. Tables 3.10 and 3.11 outline these activities for each of the two groups for the practice and operational standard-setting sessions.

Using the session observation logs, I kept a record of each participant. I recorded the time participants started and completed each activity and any events that might have related to the validity of the procedures or materials.
In the event of distracting outside noise, I provided earmuffs, which I reminded the participants to don if noise became an issue. These were used on one occasion only and for about 15 minutes by a single participant because school children were passing by in a nearby hallway.

During each session, the participants were told that they could discontinue participating at any time and still receive the payment. There was one participant who felt uncomfortable and exited the study before participating in the operational session. This individual had arrived at the room 15 minutes after the scheduled time and appeared to be very tired and busy. The participant decided to discontinue participation after the practice session because of discomfort in using the mouse and keyboard as well as physical (respiratory) discomfort which appeared to be unrelated to the task. I concluded that this participant’s state was due to conditions prior to the session and not to the session itself. This person’s random group selection card, which was for the treatment group, was placed back in the group assignment bag for a future participant to select. No other unexpected events occurred during the sessions.

After each session concluded, I saved the Excel files to a master folder and backed them up on a server. After the completion of the final session, I copied each participant’s ratings to a master Excel sheet for subsequent analysis in SAS.
Table 3.10

Practice Session Control and Treatment Activities

<table>
<thead>
<tr>
<th>Group</th>
<th>Round 1</th>
<th>Between-round activities</th>
<th>Round 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Control</td>
<td>Judges rated 10 practice-set items.</td>
<td>Judges were asked to select three items they believed they could more closely examine.</td>
<td>Judges rated the 10 practice-set items a second time.</td>
</tr>
<tr>
<td></td>
<td>Judges were provided with the answer key to each item.</td>
<td>Judges were shown the items they selected and then asked to closely examine each and to type comments to help them to consolidate their thinking during the Round 2 rating.</td>
<td>Judges were provided with the answer key to each item, the rating the judge gave the item during Round 1, and the comments the judge wrote during the between-round activity.</td>
</tr>
<tr>
<td>Treatment</td>
<td></td>
<td>Judges were presented with the three items they were most inconsistent in rating, along with a description of the magnitude and direction of the inconsistency on each item.</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Judges were asked to examine the item and to type comments to help them consolidate their thinking during the Round 2 rating</td>
<td></td>
</tr>
</tbody>
</table>

*Note.* During the rounds, the answer keys and the item ratings the judge provided for each item were conditionally formatted to appear after the judge entered a rating during that round. Judges were instructed that they could change their ratings at any time within the round.
Table 3.11
Operational Session Control and Treatment Activities

<table>
<thead>
<tr>
<th>Group</th>
<th>Round 1 Between-round activities</th>
<th>Round 2 Between-round activities</th>
<th>Round 3 Between-round activities</th>
</tr>
</thead>
<tbody>
<tr>
<td>Control</td>
<td>Judges were asked to select 12 items they believed they could more closely examine.</td>
<td>Judges were provided with the answer key to each item, the rating the judge gave the item during Round 1, and the comments the judge wrote during the between-round activity.</td>
<td>Judges rated the 50 items a third time.</td>
</tr>
<tr>
<td></td>
<td>Judges were shown the items they selected and then asked to closely examine each and to type comments to help them to consolidate their thinking during Round 2.</td>
<td>Judges were shown the items they selected and then asked to closely examine each and to type comments to help them to consolidate their thinking during Round 3.</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Judges were presented with the 12 items they were most inconsistent in rating, along with a description of the magnitude and direction of the inconsistency on each item.</td>
<td>Judges were presented with the 12 items they were most inconsistent in rating, along with a description of the magnitude and direction of the inconsistency on each item.</td>
<td>Judges were provided with the answer key to each item, the rating the judge gave the item during Round 2, and the comments the judge wrote during the previous between-round activity.</td>
</tr>
<tr>
<td></td>
<td>Judges were asked to examine the item and to type comments to help them consolidate their thinking during Round 2.</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Treatment</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Note. During the rounds, the answer keys and the item ratings the judge provided for each item were conditionally formatted to appear after the judge entered a rating during that round. Judges were instructed that they could change their ratings at any time within the round.
Preparation of the Data for Analyses

Raw data. To prepare the data for the analyses, I copied and pasted each participant’s intrajudge-consistency indexes ($C_{jr}$; one per round) and Angoff ratings (50 per round) into a master Excel file. I verified that the data were accurately copied by conducting a second copy-and-paste procedure on a separate occasion and verifying that the two sets were identical. I then imported these data, to six decimal places, from the master Excel file into a SAS data file. I also imported participants’ background-questionnaire responses from the SurveyMonkey Excel file into a SAS data file and merged these data with the $C_{jr}$ and Angoff-rating data.

Participant-background covariate variables. I generated three participant-background covariate variables: content expertise, assessment expertise, and secondary experience. Each participant’s score on the covariate was based on responses on the background questionnaire (in Appendix I). Possible content-expertise scores ranged from 0.0 to 2.2, and were calculated as the mean of five responses, one from Question 15 and four from Question 19 (responses on Question 15 were coded as $0 = \text{none or less than 1 year}$, $1 = \text{between 1 and 3 years}$, $2 = \text{between 3 and 5 years}$, and $3 = \text{5 or more years}$; responses on Question 19 were coded as $0 = \text{none}$, $1 = \text{some}$, and $2 = \text{a lot}$). Possible assessment-expertise scores ranged from 0.0 to 2.0, and were calculated as the mean of the seven items in Question 17 (where for each, $0 = \text{none}$, $1 = \text{some}$, and $2 = \text{a lot}$). Secondary-experience scores ranged from 0 to 3 (in integers), and were operationalized as the response to Question 3 (where $0 = \text{none or less than 1 year}$, $1 = \text{between 1 and 3 years}$, $2 = \text{between 3 and 5 years}$, and $3 = \text{5 or more years}$).
Analysis for Research Question 1

The first research question asked whether there would be a difference in gains in intrajudge consistency, across the three rounds of rating, between the group of judges receiving intrajudge-consistency feedback and the group of judges asked to select and reexamine items on their own. Thus, the outcome variable was the intrajudge-consistency index of each judge, and the parameter under investigation was the slope of the Time-by-Group interaction, where Time represented a quantification of the round and Group was a categorical variable for experimental group.

Because ordinary-least-squares estimation in repeated-measures models assumes sphericity (a compound-symmetric error covariance matrix) and homogeneity of error variances across time points, I used a multilevel-model-for-change framework (Singer & Willett, 2003), which uses maximum-likelihood estimation methods and allows for specification of the within-subjects covariance structures (relaxing these statistical assumptions). For model estimation, I used the MIXED procedure (PROC MIXED) in SAS 9.3, with round specified as the Level 1 (within-subjects) unit and judge specified as the Level 2 unit. Because my study included a somewhat small number of Level 2 units (36 judges) and it included complex within-subjects covariance structures, I used the Kenward-Roger denominator degrees of freedom method, as recommended by Littell, Milliken, Stroup, Wolfinger, and Schabenberger (2006); this method is more conservative against Type I error rate than the default between-within method and is more appropriate when data slightly deviate from normality.

Statistical assumptions. The statistical assumptions underlying analyses of longitudinal data using PROC MIXED are linearity, normal distribution of residuals, and
independence of the Level 2 units (Singer & Willett, 2003). If the trend of change over
time appears to deviate from linearity, either a transformation of the time variable should
be considered or a model allowing nonlinear relationships among variables (such as
PROC NLMIXED) should be used (Singer & Willett, 2003). Homogeneity of variance
between time variables can also be assumed depending on the specified covariance
structure, which I deal with in the model comparisons section below.

In this study, the case for independence among subjects can be made based on the
design and the data from session observation logs. An investigation of normality can be
conducted by first screening the data to determine whether the outcome variable should
be transformed, and then examining the normality of the residuals after identifying the
optimal multilevel model (which I describe below). The assumption of normality of
residuals may take a bivariate form for the Level 2 residuals if both the intercept and
slope are included as outcomes in the final optimal model.

**Independence of the Level 2 units.** Among the 23 standard-setting sessions, 10 were
attended by a single judge, suggesting these judges made their ratings independent of the
ratings of other judges. Following the facilitator protocol, I instructed judges in the 13
other sessions to avoid discussing any features of the study with any other judges during
the operational sessions. The data in the session observation logs provided no evidence
that these judges discussed items with each other or that any particular judge’s ratings
affected another judge’s ratings. Thus, based on the design, I believed the assumption of
independence of the Level 2 units (judges) had been met.
Data screening. Prior to conducting the analysis, I screened the raw data, examining the descriptive statistics, linearity, reliability, normality, and outliers. This involved the following steps:

1. I calculated the descriptive statistics of the $C_{jr}$ estimates for the total group of 36 judges and for each of the two experimental groups at each of the three rounds. The total-group data are what are used in the unconditional growth model (when time is included in the model, but experimental group is not), and help to identify whether a linear or nonlinear trend would be appropriate. To examine this trend over time, I plotted the observed total-group data over each round. This helped to identify whether a linear-trend would suffice or whether a transformation of time would be appropriate. I also examined Cronbach’s alpha reliability for each of the two groups, across all three rounds, and for the total group.

2. To determine whether the consistency indexes needed to be transformed in order to make them more normally distributed, I conducted Shapiro-Wilk significance tests, examined the confidence limits of kurtosis and skewness, and plotted the normal-probability plots. I also checked for outliers in the box-plots of each round for each group. After I identified the optimal model for addressing Research Question 1 (described in the model comparison section), I examined the assumption of normality of the residuals.

3. With the participant-background covariate data, I examined the raw item responses and calculated the descriptive statistics of the covariate variable scores. This was to ensure the composite scores had been calculated correctly and to flag any missing or unusual scores. With the two variables (content expertise and
assessment expertise) that were calculated based on the mean responses across items, I also examined Cronbach’s alpha reliability.

4. Following the advice of Shadish, Cook, and Campbell (2002), I examined the correlations among the covariates and the outcome variables (the consistency index of each round) used for addressing Research Question 1. I calculated Pearson correlation coefficients and inspected the scatter plots (to identify any non-linear trends not captured by the correlation coefficients) of 12 pairs of variables. Three of these correlations were between covariates (i.e., content expertise with assessment expertise, content expertise with secondary experience, and assessment expertise with secondary experience) and nine were between each covariate and each round’s consistency index. Correlations among covariates helped me to determine whether any covariates were redundant with each other, which is important information for determining whether their inclusion in any model would result in a waste of degrees of freedom (Shadish et al., 2002). Correlations between each covariate and each round of consistency index helped me to determine whether there appeared to be any patterns in the change in consistency indexes across rounds that might be due to the background experience or expertise—this information would further justify investigation into whether these covariates should be included in the model, which I examined in the subsequent model-comparison procedures.

Model selection. Before answering Research Question 1, I needed to identify an optimal multilevel model that maximized parsimony with negligible expense to explanatory power. To identify this optimal model, I conducted two stages of model
comparison, which I describe below, and followed Singer and Willett’s (2003) suggestions for model building.

**Criteria for model selection.** In balancing simplicity and explanatory power to identify the optimal model, I used the following criteria: The model had to (a) adequately estimate the trend over time, (b) include participant-background covariates, if any, that significantly contribute to model fit, (c) include the simplest and best-fitting within-subjects covariance matrix structure, and (d) include random effects (such as intercepts and intercepts-and-slopes as outcomes), in addition to a within-subjects covariance matrix structure, if they explained significantly more variability than the best fixed-effects model with a within-subjects covariance structure. I addressed the first two criteria in the first stage of model comparisons and the last two criteria in the second stage of comparisons.

**First stage of model comparisons.** In the first stage, I examined the increase in explanatory power (more accurately, the reduction in error variance) as I included predictors and participant-background covariates. Table 3.12 shows the specifications of the models I compared. For these model comparisons, I calculated the improvements in explained variability as estimated from the pseudo-$R^2$ calculations, using equations in Singer and Willett (2003), and the goodness-of-fit statistics.

Pseudo-$R^2$ estimates provide an estimate of the proportional reduction in residual variance after adding predictors and covariates to the model (Singer & Willett, 2003). I used the pseudo-$R^2$ estimates to examine the reduction in Level-1 error variance after modeling time (when comparing Model B to Model A), and the reduction in Level-2 error variance after including experimental group and the covariates in the model (to
examine Models C and D). The Level-1 error variance was the unexplained within-judge variability (interpretable in the same way as in traditional regression). The Level-2 error variance was made up of two components: the intercept variance, $\tau_{00}$, and the slope variance $\tau_{11}$, where $\tau_{00}$ estimated the error in predicting judges’ Round 1 consistency indexes, and $\tau_{11}$ estimated the error in predicting the judges’ changes in consistency indexes from round to round.

To compare models using the goodness-of-fit statistics, I examined the deviance (-2LL) and the Akaike information criterion (AIC) statistics. Specifically, I conducted chi-square significance tests on improvement in -2LL fit indexes and verified these results with the AIC estimates. I used full maximum likelihood (FML) estimation, which permits comparison of fit statistics in models with different fixed parameters.

In multilevel models, the intercept-slope covariance component, $\tau_{01}$, can also be estimated. This covariance estimate explains how much of the unexplained variability in judges’ change over round is related to judges’ Round 1 ratings. For interpreting whether this covariance was meaningful in any of the models under comparison, I transformed this covariance into the correlation estimate, calculated as $\frac{\tau_{01}}{\sqrt{\tau_{00} + \tau_{11}}}$.

With Models A through D, the Level 1 residual was assumed to be normally distributed around zero, with an unknown population variance $\sigma^2_\varepsilon$; that is, $\varepsilon_{jr} \sim N(0, \sigma^2_\varepsilon)$. The unexplained variability of the Level 2 parameters was assumed to be bivariate-normal around a mean of zero; that is, $\begin{bmatrix} \mu_{0j} \\ \mu_{1j} \end{bmatrix} \sim \left( \begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \tau_{00} & \tau_{01} \\ \tau_{10} & \tau_{11} \end{bmatrix} \right)$, with the terms for the slope and intercept-slope interaction reducing to zero for Model A.
Table 3.12  
*Specifications of Models in the First Stage of Model Comparisons*

<table>
<thead>
<tr>
<th>Model</th>
<th>Level 1</th>
<th>Level 2</th>
<th>Combined</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>$C_{jr} = \beta_{0j} + \epsilon_{jr}$</td>
<td>$\beta_{0j} = \gamma_{00} + \mu_{0j}$</td>
<td>$C_{jr} = \gamma_{00} + \mu_{0j} + \epsilon_{jr}$</td>
</tr>
<tr>
<td></td>
<td>$\beta_{1j} = \gamma_{10} + \mu_{1j}$</td>
<td></td>
<td></td>
</tr>
<tr>
<td>B</td>
<td>$C_{jr} = \beta_{0j} + \beta_{1j}(Time)<em>{jr} + \epsilon</em>{jr}$</td>
<td>$\beta_{0j} = \gamma_{00} + \mu_{0j}$</td>
<td>$C_{jr} = \gamma_{00} + \gamma_{10}(Time)<em>{jr} + \mu</em>{0j} + \mu_{1j}(Time)<em>{jr} + \epsilon</em>{jr}$</td>
</tr>
<tr>
<td></td>
<td>$\beta_{1j} = \gamma_{10} + \mu_{1j}$</td>
<td></td>
<td></td>
</tr>
<tr>
<td>C</td>
<td>$C_{jr} = \beta_{0j} + \beta_{1j}(Time)<em>{jr} + \epsilon</em>{jr}$</td>
<td>$\beta_{0j} = \gamma_{00} + \gamma_{01}(Group)<em>{j} + \mu</em>{0j}$</td>
<td>$C_{jr} = \gamma_{00} + \gamma_{01}(Group)<em>{j} + \gamma</em>{10}(Time)<em>{jr}$ + $\gamma</em>{11}(Time*Group)<em>{jr} + \mu</em>{0j} + \mu_{1j}(Time)<em>{jr} + \epsilon</em>{jr}$</td>
</tr>
<tr>
<td></td>
<td>$\beta_{1j} = \gamma_{10} + \gamma_{11}(Group)<em>{j} + \mu</em>{1j}$</td>
<td></td>
<td></td>
</tr>
<tr>
<td>D</td>
<td>$C_{jr} = \beta_{0j} + \beta_{1j}(Time)<em>{jr} + \epsilon</em>{jr}$</td>
<td>$\beta_{0j} = \gamma_{00} + \gamma_{01}(Group)<em>{j} + \mu</em>{0j}$</td>
<td>$C_{jr} = \gamma_{00} + \gamma_{01}(Group)<em>{j} + \gamma</em>{10}(Time)<em>{jr}$ + $\gamma</em>{11}(Time<em>Group)<em>{jr} + \gamma</em>{12}(Time</em>Covariate)<em>{jr} + \mu</em>{0j} + \mu_{1j}(Time)<em>{jr} + \epsilon</em>{jr}$</td>
</tr>
<tr>
<td></td>
<td>$\beta_{1j} = \gamma_{10} + \gamma_{11}(Group)_{j}$</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>+ $\gamma_{12}(Covariate)<em>{j} + \mu</em>{1j}$</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

*Note.* The outcome variable, $C_{jr}$, was the intrajudge-consistency index of the $j$th judge in the $r$th round.
Unconditional means model. Model A, the unconditional means model, provided the Level 1 and Level 2 estimates of variability in the model. I used these to calculate the intraclass correlation coefficient and to establish whether there was empirical evidence to support further investigation into models with predictors. This model also provided a baseline for examining whether Model B, with Time as a Level 1 predictor, better explained the Level 1 variability.

Unconditional growth model. To address my first criterion for an optimal model, I examined two versions of Model B: Models B1 and B2. In Model B1, I scaled Time linearly, coding Round 1 as 0.00, Round 2 as 1.00, and Round 3 as 2.00. In Model B2, I transformed Time onto a logarithmic scale, coding Round 1 as 0.00, Round 2 as 0.69, and Round 3 as 1.10. Investigating the fit of these two models allowed me to compare the linear-trend with a logarithmic-trend model in order to identify which one to use as the unconditional growth model.

Whereas other nonlinear transformations and models were possible, the logarithmic-trend resembled the observed data and was consistent with existing theory that feedback has a stronger effect the earlier it is provided in the standard-setting process (Cizek & Bunch, 2007; Reckase, 2001). I did not investigate a quadratic or reverse-exponential model because my dataset had only three time points, and these models would have over-parameterized the model (as explained in Holt, 2008).

Experimental group as a predictor. Model C included Group as a Level 2 main effect and Time-by-Group as a cross-level interaction effect. Group was a dichotomous variable for experimental group status, where the control-group judges were coded as 0 and the treatment-group judges were coded as 1.
I compared this model with the unconditional growth model (Model B), to identify whether inclusion of this predictor better explained Level 2 variability of the intercepts and slopes. That is, I examined the extent to which experimental group status explained variation among judges in their (a) initial (Round 1) consistency indexes and (b) changes in consistency indexes from round to round.

*Participant-background covariate models.* To address the second criterion in model selection, I examined the three models with the participant-background covariates. Thus, there were three versions of Model D: Model D1 included the content-expertise background covariate, Model D2 included the assessment-expertise covariate, and Model D3 included the secondary-teaching-experience covariate. (Operationalization of these covariates is described above in the section explaining how the data were prepared for analyses.)

I examined these covariate models to identify whether any of these participant-background variables added precision to the model and to rule out alternative, possibly confounding, explanations for the gains in $C_{jr}$. If any covariate models showed a significant Time-by-Covariate cross-level effect or explained more of the Level 2 variance than Model C, their covariates were to be included as control-variables in the final optimal model. Covariates that did not improve fit were to be excluded from further analysis.

*Second stage of comparisons.* After establishing whether to include a linear- or logarithmic-trend model and identifying which, if any, covariates should be included in estimating change over time, I focused on specifying the residual covariance structure
and any remaining unexplained Level 2 variability as specified by intercepts-as-outcomes and intercepts-and-slopes-as-outcomes models.

**Covariance structure specifications.** Whereas the models in the first stage of comparisons included both Level 1 and Level 2 unexplained variances, their counterparts modeled with a residual covariance structure facilitate interpretation by combining the Level 1 and Level 2 unexplained variances into a single Level 1 residual, \( r_{jr} \) (Singer & Willett, 2003). This also allows for investigation into alternative covariance structures besides the one imposed by those in the models in the first stage (in which the variance estimates are dependent on time; Singer & Willett, 2003). Using this approach, Model C, for example, would be specified in this way:

\[
\begin{align*}
\text{Level 1:} & \quad C_{jr} = \beta_{0j} + \beta_{1j}(\text{Time})_{jr} + r_{jr} \\
\text{Level 2:} & \quad \beta_{0j} = \gamma_{00} + \gamma_{01}(\text{Group})_j \\
& \quad \beta_{1j} = \gamma_{10} + \gamma_{11}(\text{Group})_j \\
\text{Combined:} & \quad C_{jr} = \gamma_{00} + \gamma_{01}(\text{Group})_j + \gamma_{10}(\text{Time})_{jr} + \gamma_{11}(\text{Time*Group})_{jr} + r_{jr}
\end{align*}
\]

With this specification, the residual is assumed to be normally distributed around zero, with a covariance structure \( \Sigma \); that is, \( r_{jr} \sim N(0, \Sigma) \).

With these models, I used restricted-maximum likelihood (REML) estimation because it is more appropriate for modeling residual covariance and provides less biased estimates of the fixed effects when sample sizes are small (Fitzmaurice, Laird, & Ware, 2011; Littell et al., 2006; Singer & Willett, 2003). In comparing models, I used the same deviance comparison procedures that I used during the first stage of model comparisons.

**Covariance structure model comparisons.** To address the third criterion, I compared covariance-structure models for parsimony and goodness of fit. The covariance structures
I examined (in order of most to least complex) included unstructured (UN), heterogeneous first-order autoregressive [ARH(1)], Toeplitz (TOEP), heterogeneous compound symmetry (CSH), first-order autoregressive [AR(1)], and compound symmetry (CS) structures. Although other covariance structures are possible, I selected these because Singer and Willett (2003) suggest they are the most useful structures in longitudinal repeated-measures modeling. The TOEP, AR(1), and CS models assume homogeneity of variance of the residuals across occasions (rounds, or Time in the model), but differ in the constraints they impose on the covariances among the occasions.

These covariance-structure models (up to this point) did not specify any Level 2 random effects, and were, therefore, fixed-effects models. Because it was possible, however, for there to still be random variation among judges in their intercepts and slopes after imposing the covariance structure, I reserved final judgment about the best covariance structure until after examining random effects models, as suggested by Singer (1998).

*Intercepts- and intercepts-and-slopes-as-outcomes models.* To address the last criterion, I compared, against the best fixed-effects covariance structure model, the models with intercepts as outcomes and intercepts-and-slopes as outcomes. I also reexamined the covariance structures with these models (when they successfully converged), because, according to Singer (1998), simpler covariance structures may show improved fit after random effects have been included in the model.

*The optimal model.* The model with the best fit and parsimony, as defined by the four criteria, served as the optimal model for answering Research Question 1. Before
using this model to answer Research Question 1, I examined the normality of the residuals using the Shapiro-Wilk significance test and a Q-Q plot of the residuals.

**Analysis for Research Questions 2 and 3**

To address the last two research questions, I used a G-theory framework. The variable under investigation was the Angoff ratings that judges provided on each item. I carried out the following steps:

1. To aid the interpretation of the generalizability study, I calculated the descriptive statistics of the two experimental groups’ recommended cutscores in each of the three rounds. This involved first summing each judge’s set of Angoff ratings (in each round) to arrive at that judge’s recommended cutscore (for each round).

2. For Research Question 2, I conducted a G study, and specified the composition of the variability of the observed item-level Angoff ratings, $\sigma_{\text{obsij}}^2$, as

   $$\sigma_{\text{obsij}}^2 = \sigma_i^2 + \sigma_j^2 + \sigma_{ij,e}^2,$$

   where $\sigma_i^2$ = the item variance, $\sigma_j^2$ = the judge variance, and $\sigma_{ij,e}^2$ = the item-by-judge interaction variance, which is conflated with unexplained error variance. Items were the object of measurement.

3. I used SAS PROC MIXED, with Type I sums-of-squares estimation, to calculate the estimated variance components. To verify that the estimation was accurate, I ran a general-linear-model procedure (PROC GLM in SAS) and cross-checked the estimated variance components based on Excel calculations on the estimated PROC GLM sums of squares. Additionally, I verified that my Excel calculations were correct by imputing sums of squares presented in Brennan (2001) and confirming that my variance components estimates were identical to those Brennan reported.
4. In SAS, I calculated, for each of the two experimental groups for each round of rating, the percent of variance explained by the items, the judges, and item-by-judge interaction. I verified the accuracy of the calculations using Excel.

5. I calculated the dependability index based on the variance components and equations in Brennan (2001, Equations 2.32 and 2.41).

6. I calculated the standard error of the cutscore for generalizing over future sets of items, $\sigma_{X}$, by taking the square root of Brennan and Lockwood’s (1980)

Equation 8 estimate of the variance, $\hat{\sigma}_{X}^2$, where $\hat{\sigma}_{X}^2 = \frac{\hat{\sigma}_i^2}{n_i} + \frac{\hat{\sigma}_j^2}{n_j} + \frac{\hat{\sigma}_{ij}^2}{n_in_j}$. This treats items and judges as random components. I calculated the standard error of the cutscore for the specific set of items, $\hat{\sigma}_{Xl}^2$, using the square root of Brennan and Lockwood’s Equation 10 estimate of the variance, $\hat{\sigma}_{Xl}^2$, where $\hat{\sigma}_{Xl}^2 = \frac{\hat{\sigma}_j^2}{n_j} + \frac{\hat{\sigma}_{ij}^2}{n_in_j}$. This treats judges as random and items as fixed (also see Raymond & Reid, 2001).

7. For Research Question 3, I conducted a D study on both experimental groups’ Round 1 and Round 3 Angoff ratings. This involved varying the proposed number of judges from 1 to 20 and estimating the standard error of the cutscore when the item and judge components are specified as random. I transformed the standard error estimates onto the raw scale, using methods described in Hurtz and Hertz (1999), to identify the minimum number of judges that would be required (for each of the two experimental groups) for a standard error of the cutscore that would be equivalent to just less than 2.50 items. To examine how many judges would be needed to meet the .80 dependability-index criterion, I calculated the dependability indexes for each number-of-judge scenario.
CHAPTER 4
RESULTS
This chapter begins with a description of the observed intrajudge-consistency index data, which I use for answering Research Question 1. This description includes the descriptive statistics, a visual representation of the trend over time, reliability estimates, and information about the normality of the raw consistency-index data in each round. I also present the descriptive statistics, reliability, and correlation coefficients of the participant-background covariate variables. Following these descriptions, I report the results of the model comparisons and describe the resulting optimal model, including its parameters, which I use for answering Research Question 1. In the latter part of the chapter, I report the G-theory results used for addressing Research Questions 2 and 3. I start with a report of the descriptive statistics of the cutscores and follow with the G-study and D-study results.

There were no missing data. That is, every participant provided an Angoff rating on every item in every round and responded to every question asked of them in the online background questionnaire. Thus, all participants’ intrajudge-consistency indexes and participant-background covariate scores were included in the analysis for Research Question 1; furthermore, the fully-crossed Angoff-rating data facilitated the G-theory analyses used for addressing the remaining two research questions.

Intrajudge Consistency Indexes

Descriptive statistics. Descriptive statistics of the intrajudge-consistency indexes of the total group of judges and of each of the two experimental groups are shown in Table 4.1. The consistency indexes of the total group improved by about .03 points from Round
1 to Round 3, which was slightly lower than the standard deviations within each round (about .04), suggesting moderate growth in the aggregate index values of the judges.

Figure 4.1 displays the plot of the means across rounds for the total group. Although the 95% confidence limit bands suggested a linear trend might be acceptable for the unconditional growth model, the shape indicated that a logarithmic trend would likely be more precise.

Table 4.1

<table>
<thead>
<tr>
<th>Group</th>
<th>Round</th>
<th>M</th>
<th>SD</th>
<th>Low</th>
<th>Q1</th>
<th>Median</th>
<th>Q3</th>
<th>High</th>
</tr>
</thead>
<tbody>
<tr>
<td>Control</td>
<td>1</td>
<td>.8163</td>
<td>.0424</td>
<td>.7221</td>
<td>.7986</td>
<td>.8155</td>
<td>.8331</td>
<td>.9099</td>
</tr>
<tr>
<td>Total</td>
<td>2</td>
<td>.8400</td>
<td>.0440</td>
<td>.7547</td>
<td>.8116</td>
<td>.8409</td>
<td>.8730</td>
<td>.9297</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>.8473</td>
<td>.0476</td>
<td>.7389</td>
<td>.8225</td>
<td>.8465</td>
<td>.8801</td>
<td>.9334</td>
</tr>
<tr>
<td>Control</td>
<td>1</td>
<td>.8240</td>
<td>.0463</td>
<td>.7221</td>
<td>.7990</td>
<td>.8226</td>
<td>.8460</td>
<td>.9067</td>
</tr>
<tr>
<td>Treatment</td>
<td>2</td>
<td>.8234</td>
<td>.0437</td>
<td>.7547</td>
<td>.7876</td>
<td>.8232</td>
<td>.8499</td>
<td>.9052</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>.8214</td>
<td>.0448</td>
<td>.7389</td>
<td>.7834</td>
<td>.8225</td>
<td>.8491</td>
<td>.9024</td>
</tr>
<tr>
<td>Control</td>
<td>1</td>
<td>.8086</td>
<td>.0378</td>
<td>.7334</td>
<td>.7964</td>
<td>.8082</td>
<td>.8223</td>
<td>.9099</td>
</tr>
<tr>
<td>Treatment</td>
<td>2</td>
<td>.8567</td>
<td>.0385</td>
<td>.7773</td>
<td>.8351</td>
<td>.8561</td>
<td>.8890</td>
<td>.9297</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>.8731</td>
<td>.0352</td>
<td>.8304</td>
<td>.8438</td>
<td>.8616</td>
<td>.9051</td>
<td>.9334</td>
</tr>
</tbody>
</table>

Note. The total group included all 36 judges, with 18 from each experimental group. Effect sizes for treatment-control comparisons were the following: Round 1 Hedge’s $g = -0.36$, Round 2 Hedge’s $g = 0.81$; Round 3 Hedge’s $g = 1.28$. Q1 = first quartile; Q3 = third quartile.

In Round 1, before any between-round activities, the control group’s mean consistency index was slightly higher than that of the treatment group. The mean consistency index of the control group stayed about the same across the three rounds whereas that of the treatment group increased from Round 1 to Round 3 by around .06
points. The median indexes roughly followed the same pattern, as did the indexes at the first and third quartile and high locations; at the low location, the control group showed an increase. For both groups, the standard deviations were less than .05 points, with the treatment group showing less variability across judges. These descriptive statistics along with the negative between-group Hedge’s g effect size in Round 1 and the strong positive effect sizes in Rounds 2 and 3 supported further investigation into whether the two groups statistically differed in their slopes.

Figure 4.1. Plot of the mean observed consistency index per round across all 36 judges. Error bars represent the 95% confidence interval of each mean, calculated as ±2.03 x (standard error of the mean), where the 2.03 multiplier is based on the t-distribution with 35 degrees of freedom.

**Reliability of the consistency indexes.** Across the three rounds, the Cronbach’s alpha reliability of the consistency-index estimates was .96 for the control and .92 for the treatment group. Thus, participants with relatively high intrajudge consistency in one round tended to also have relatively high intrajudge consistency in the other two rounds.
Reliability for the entire group was .89. This lower estimate for the total group was not surprising because if participants in one group (in the treatment group, for example) improved in their consistency while those in the other experimental group did not, reliability would likely be lower.

**Normality and outlier detection.** The results, reported in Table 4.2, provided evidence of a minor deviation from normality. The Shapiro-Wilk tests indicated that only one of the six datasets, Round 3 of the treatment group, likely deviated from normality, $W = .89, p = .046$. The normal probability plots (Figure 4.2) revealed that with this dataset, the participants at the extremes tended to have slightly higher consistency indexes than would be expected if the data were perfectly normal. This may have been due to a slight ceiling effect as a result of the treatment. Kurtosis and skewness of this dataset were not significant; thus, whereas some deviation from normality was present, it appeared not to be severe enough to warrant data transformation.

<table>
<thead>
<tr>
<th>Group</th>
<th>Round</th>
<th>Kurtosis Estimate</th>
<th>LB</th>
<th>UB</th>
<th>Skewness Estimate</th>
<th>LB</th>
<th>UB</th>
<th>Shapiro-Wilk test</th>
</tr>
</thead>
<tbody>
<tr>
<td>Control</td>
<td>1</td>
<td>0.45</td>
<td>-1.99</td>
<td>2.88</td>
<td>-0.23</td>
<td>-1.45</td>
<td>0.99</td>
<td>0.97, .800</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>-0.54</td>
<td>-2.97</td>
<td>1.90</td>
<td>0.20</td>
<td>-1.02</td>
<td>1.42</td>
<td>0.96, .645</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>-0.57</td>
<td>-3.00</td>
<td>1.87</td>
<td>0.08</td>
<td>-1.14</td>
<td>1.30</td>
<td>0.98, .972</td>
</tr>
<tr>
<td>Treatment</td>
<td>1</td>
<td>2.59</td>
<td>0.16</td>
<td>5.03</td>
<td>0.63</td>
<td>-0.59</td>
<td>1.85</td>
<td>0.93, .191</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>-0.11</td>
<td>-2.55</td>
<td>2.33</td>
<td>-0.20</td>
<td>-1.42</td>
<td>1.01</td>
<td>0.98, .938</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>-1.07</td>
<td>-3.51</td>
<td>1.37</td>
<td>0.60</td>
<td>-0.62</td>
<td>1.82</td>
<td>0.89, .046</td>
</tr>
</tbody>
</table>

**Note.** Confidence intervals of the kurtosis and skewness estimates were calculated by multiplying 2.11 (the t-value at the 95% confidence limit with 18 judges) by the standard error (SE) of each estimate, $SE_{kurt} = \sqrt{24/18}$ and $SE_{skew} = \sqrt{6/18}$. LB = lower bound 95% confidence limit; UB = upper bound 95% confidence limit.
No dataset was significantly platykurtic or skewed. However, with the treatment group’s Round 1 data, there was significant leptokurtosis (kurtosis = 2.59) even though the Shapiro-Wilk test revealed no evidence of deviation from normality ($W = .93$, $p = .191$). The probability plot of this dataset (in the top right panel of Figure 4.2) revealed that the highest index value looked like an outlier. After checking to make sure
this individual’s index was correctly calculated and entered, and after noticing that this
high consistency index (.9099) was at nearly the same location as its counterpart (.9067)
in the control group’s Round 1 ratings, I saw no reason to take action on the outlier.
Probability plots and box-plots for the remaining datasets provided no evidence of any
other obvious outliers or patterns of non-normality.

Because the evidence was weak that the data violated the assumption of normality
and that there was little disturbance from outliers, I judged the raw consistency indexes to
be adequate for the multilevel-model analysis. Interpretation of the results, however,
should be made with caution given the slight deviation from normality of the treatment
group’s last round of consistency-index data. I conducted further analysis of the
normality assumption after calculating the residuals of the final optimal model, which is
presented in the model comparisons section below.

Participant-background Covariates

Descriptive statistics. The descriptive statistics of participant-background covariate
scores are presented in Table 4.3. For each of these, a score of zero represented no
experience or expertise, which facilitated potential subsequent interpretation of the
intercepts in the multilevel models. The standard deviations indicated there was
variability among the participants in each of the variables, which provided support for
further investigation into whether these covariates would contribute to changes in
consistency-index estimates across rounds. All were somewhat platykurtic and slightly
positively skewed, suggesting conclusions about their relationships with other variables
should be interpreted with caution.
Table 4.3

Descriptive Statistics of the Participant-background Covariate Scores

<table>
<thead>
<tr>
<th>Covariate</th>
<th>$M$</th>
<th>$SD$</th>
<th>Low</th>
<th>Median</th>
<th>High</th>
<th>Kurtosis</th>
<th>Skew</th>
</tr>
</thead>
<tbody>
<tr>
<td>Content expertise</td>
<td>0.71</td>
<td>0.67</td>
<td>0.00</td>
<td>0.50</td>
<td>2.20</td>
<td>-0.77</td>
<td>0.68</td>
</tr>
<tr>
<td>Assessment expertise</td>
<td>0.92</td>
<td>0.52</td>
<td>0.00</td>
<td>0.86</td>
<td>2.00</td>
<td>-0.52</td>
<td>0.48</td>
</tr>
<tr>
<td>Secondary experience</td>
<td>1.44</td>
<td>1.33</td>
<td>0.00</td>
<td>1.00</td>
<td>3.00</td>
<td>-1.77</td>
<td>0.08</td>
</tr>
</tbody>
</table>

Note. Scores were calculated based on responses on the background questionnaire (provided in Appendix I). Content expertise was the mean response on five questionnaire items (in Questions 15 and 19), Assessment expertise was the mean response on seven questionnaire items (in Question 17), and Secondary experience was the response on a single item (Question 3).

Reliability of the covariates. With the two covariates (content expertise and assessment expertise) that were calculated as the mean of the participants’ responses on multiple background questionnaire items, the reliability was high. Cronbach’s alpha of the set of five items making up the content expertise score was .87; for the set of seven items making up the assessment expertise score, it was .89. It was not possible to calculate the reliability of the secondary experience covariate because this score was based on a single questionnaire item.

Correlations among covariates. The correlations among covariates, summarized in Table 4.4, indicated that there were no strong relationships among these variables. In my examination of the scatter plots, I found no evidence of nonlinear bivariate patterns. Even though two out of the three correlations were significant, their strength was too weak to justify removal of any of the covariates before examining them in the model comparisons.

Correlations between each covariate and each round’s consistency index. The correlations between each covariate and the outcome-measure variable of each round (for Research Question 1) are summarized in Table 4.4. (In my examination of the scatter
plots of every pair of variables, I found no evidence of nonlinear bivariate patterns.) The Pearson correlations revealed no statistically significant relationships between each covariate and each round of rating, but it appeared that both content expertise and assessment expertise may have become more positively related to consistency indexes as the rounds progressed. These findings supported further analysis (which I describe in the model-comparison section below) into whether each covariate was significantly related to the change in consistency indexes across rounds.

Table 4.4  
*Correlations between Participant-background Covariates and Consistency Indexes*

<table>
<thead>
<tr>
<th>Covariate</th>
<th>Content expertise</th>
<th>Assessment expertise</th>
<th>Secondary experience</th>
<th>C_{j1}</th>
<th>C_{j2}</th>
<th>C_{j2}</th>
</tr>
</thead>
<tbody>
<tr>
<td>Content expertise</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>Assessment expertise</td>
<td>.09</td>
<td>(.357)</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>Secondary experience</td>
<td>-.23</td>
<td>-.20</td>
<td>(.018)</td>
<td>(.042)</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>C_{j1}</td>
<td>-.15</td>
<td>-.10</td>
<td>.22</td>
<td>(.386)</td>
<td>(.558)</td>
<td>(.190)</td>
</tr>
<tr>
<td>C_{j2}</td>
<td>.12</td>
<td>(.481)</td>
<td>.22</td>
<td>.74</td>
<td>&lt;.001</td>
<td>—</td>
</tr>
<tr>
<td>C_{j3}</td>
<td>.27</td>
<td>(.113)</td>
<td>.20</td>
<td>.51</td>
<td>.92</td>
<td>&lt;.001</td>
</tr>
</tbody>
</table>

*Note.* The *p*-values are in parentheses beneath each correlation estimate. C_{jr} = consistency index during Round r.

**Model Comparisons for Predicting Intrajudge Consistency Indexes**

To identify the best fitting and most parsimonious model, which was subsequently used for answering Research Question 1, I conducted model comparisons. In the first stage, I examined the trend over time and the effect sizes of adding predictors to the models, operationalized as improved pseudo-$R^2$ estimates. In the second stage, I
identified the optimal covariance structure and determined whether a random-effects model would more optimally specify the model.

**Unconditional means model.** The unconditional model provided a comparison model for examining whether models with predictors decreased the unexplained variability in the data. The parameter estimates of this model are shown in Table 4.5 under Model A. The intercept, estimated as .8345, was interpreted as the mean consistency index of all 36 judges across all three rounds. The estimated unexplained variance components of the between-judge ($\hat{\tau}_{00}$) and within-judge ($\hat{\sigma}_e^2$) effects when no predictors were included in the model were significant. The intraclass correlation coefficient, from the formula $\frac{\hat{\tau}_{00}}{\hat{\tau}_{00} + \hat{\sigma}_e^2}$, was $\frac{0.0013}{(0.0013 + 0.0008)} = 0.62$. Thus, it was estimated that 62% of the variability in the consistency indexes was due to differences among the judges and 37% percent was due to differences among rounds within judges.

Because there was variability within both the within-judge and between-judge units, I had empirical evidence supporting my investigation into whether the predictors and teacher-background covariates further explained the variability.

**Unconditional growth model.** To identifying which model to use as the unconditional growth model, I examined two models, one with time coded linearly and one with time coded logarithmically. Because of the shape of the trend in the observed data, which showed a diminishing slope from Round 2 to Round 3 (Figure 4.1), I suspected that a logarithmic model would better fit the data. Furthermore, a logarithmic model would be consistent with scholars’ (Cizek & Bunch, 2007; Reckase, 2001) assertions that the effect of feedback tends to be stronger the earlier it is in the standard-setting process. Both models’ intercepts can be interpreted as the mean consistency index
Table 4.5
Parameter Estimates in the First Stage of Model Comparisons

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Model A</th>
<th>Model B1</th>
<th>Model B2</th>
<th>Model C</th>
<th>Model D1</th>
<th>Model D2</th>
<th>Model D3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fixed effects</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Intercept ($\gamma_{00}$)</td>
<td>.8345***</td>
<td>.8191***</td>
<td>.8173***</td>
<td>.8243***</td>
<td>.8243***</td>
<td>.8243***</td>
<td>.8243***</td>
</tr>
<tr>
<td></td>
<td>(.0066)</td>
<td>(.0070)</td>
<td>(.0070)</td>
<td>(.0098)</td>
<td>(.0098)</td>
<td>(.0098)</td>
<td>(.0098)</td>
</tr>
<tr>
<td>Time ($\gamma_{10}$)</td>
<td>.0155***</td>
<td>.0288***</td>
<td>-.0022</td>
<td>-.0044</td>
<td>-.0079</td>
<td>-.0049</td>
<td></td>
</tr>
<tr>
<td></td>
<td>(.0037)</td>
<td>(.0067)</td>
<td>(.0060)</td>
<td>(.0066)</td>
<td>(.0086)</td>
<td>(.0074)</td>
<td></td>
</tr>
<tr>
<td>Group ($\gamma_{01}$)</td>
<td></td>
<td>-.0139</td>
<td>-.0139</td>
<td>-.0139</td>
<td>-.0139</td>
<td>-.0139</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>(.0139)</td>
<td>(.0139)</td>
<td>(.0139)</td>
<td>(.0139)</td>
<td>(.0139)</td>
<td></td>
</tr>
<tr>
<td>Time*Group ($\gamma_{11}$)</td>
<td></td>
<td>.0621***</td>
<td>.0591***</td>
<td>.0602***</td>
<td>.0621***</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>(.0085)</td>
<td>(.0092)</td>
<td>(.0087)</td>
<td>(.0086)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Time*Covariate ($\gamma_{12}$)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>.0052</td>
<td>.0073</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>(.0066)</td>
<td>(.0078)</td>
</tr>
<tr>
<td>Variance components</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Residual ($\sigma^2_e$)</td>
<td>.0008***</td>
<td>.0002***</td>
<td>.0001***</td>
<td>.0001***</td>
<td>.0001***</td>
<td>.0001***</td>
<td>.0001***</td>
</tr>
<tr>
<td></td>
<td>(.0001)</td>
<td>(.0000)</td>
<td>(.0000)</td>
<td>(.0000)</td>
<td>(.0000)</td>
<td>(.0000)</td>
<td>(.0000)</td>
</tr>
<tr>
<td>Intercept ($\tau_{00}$)</td>
<td>.0013***</td>
<td>.0016***</td>
<td>.0017***</td>
<td>.0017***</td>
<td>.0017***</td>
<td>.0017***</td>
<td>.0017***</td>
</tr>
<tr>
<td></td>
<td>(.0004)</td>
<td>(.0004)</td>
<td>(.0004)</td>
<td>(.0004)</td>
<td>(.0004)</td>
<td>(.0004)</td>
<td>(.0004)</td>
</tr>
<tr>
<td>Slope ($\tau_{11}$)</td>
<td>.0004***</td>
<td>.0015***</td>
<td>.0005***</td>
<td>.0005***</td>
<td>.0005***</td>
<td>.0005***</td>
<td>.0005***</td>
</tr>
<tr>
<td></td>
<td>(.0001)</td>
<td>(.0004)</td>
<td>(.0002)</td>
<td>(.0002)</td>
<td>(.0002)</td>
<td>(.0002)</td>
<td>(.0002)</td>
</tr>
<tr>
<td>Intercept*slope ($\tau_{01}$)</td>
<td></td>
<td>-.0002</td>
<td>-.0006</td>
<td>-.0004</td>
<td>-.0003</td>
<td>-.0003</td>
<td>-.0004</td>
</tr>
<tr>
<td></td>
<td>(.0002)</td>
<td>(.0003)</td>
<td>(.0002)</td>
<td>(.0002)</td>
<td>(.0002)</td>
<td>(.0002)</td>
<td>(.0002)</td>
</tr>
<tr>
<td>Pseudo-$R^2$ and fit indexes</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Model $R^2_{C,C}$</td>
<td>.749</td>
<td>.974</td>
<td>.984</td>
<td>.984</td>
<td>.984</td>
<td>.984</td>
<td>.984</td>
</tr>
<tr>
<td>Residual $R^2_e$</td>
<td>.810</td>
<td>.892</td>
<td>.892</td>
<td>.892</td>
<td>.892</td>
<td>.892</td>
<td>.892</td>
</tr>
<tr>
<td>Intercept $R^2_{r00}$</td>
<td></td>
<td>.028</td>
<td>.028</td>
<td>.028</td>
<td>.028</td>
<td>.028</td>
<td>.028</td>
</tr>
<tr>
<td>Slope $R^2_{r11}$</td>
<td>.653</td>
<td>.663</td>
<td>.665</td>
<td>.649</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>-2LL</td>
<td>-399.6</td>
<td>-453.8</td>
<td>-473.1</td>
<td>-507.6</td>
<td>-508.2</td>
<td>-508.5</td>
<td>-507.9</td>
</tr>
<tr>
<td>AIC</td>
<td>-393.6</td>
<td>-441.8</td>
<td>-461.1</td>
<td>-491.6</td>
<td>-490.2</td>
<td>-490.5</td>
<td>-489.9</td>
</tr>
</tbody>
</table>

*Note.* FML estimation was used for all models. Each estimate’s standard error is in parentheses beneath it. Model A was the unconditional means model. In Model B1, Time was coded linearly as 0, 1, and 2. Model B2 was the unconditional growth model, with Time coded as the natural log of 1, 2, and 3 (0.00, 0.69, and 1.10). Model C included the experimental group (where control was coded as 0 and treatment as 1). Models D1 – D3 differed in their covariates (D1 was content expertise, D2 was assessment expertise, and D3 was secondary-teaching experience). $R^2_{C,C} = \text{pseudo-}R^2$ of the model, where $C = \text{observed consistency index and } \hat{C} = \text{predicted consistency index.}$

*** $p < .001$. All other estimates were not significant at $\alpha = .05$. 

104
at Round 1 across experimental groups and regardless of participant-background covariates.

With linear time (a Level 1 predictor) added to the model, the percent of explained variability of the consistency indexes increased and the percent of unexplained Level 1 variability decreased. This was indicated by the pseudo-$R^2$ estimate of the model’s variance, $R_{C,\hat{C}}^2$, and of the improved residual variance, $R_e^2$, in Model B1 (displayed in Table 4.5). The $R_{C,\hat{C}}^2$ estimate of .974 indicated that the linear model explained about 97% of the observed variability in the consistency indexes, which is an improvement over the unconditional model’s .749 estimate. The $R_e^2$ estimate of .810 can be interpreted to mean that 81% more of the within-judge variability was explained by the model (over the unconditional means model) when linear time was included as a predictor.

The logarithmic-trend model, Model B2, explained more within-judge variability (89% over the unconditional means model) than the linear-trend model, suggesting it would be more precise. Furthermore, both models had the same number of parameters, but the logarithmic model had a lower deviance (-2LL) and AIC estimate than the linear model by 19.3 points, indicating better fit at no expense to parsimony. Because of this evidence and because it is consistent with existing theories on standard-setting feedback, I selected the logarithmic model as the unconditional growth model upon which to build subsequent models.

**Experimental group as predictor.** Model C included Group as a Level 2 predictor of the model. Because the control group was coded as 0 and the treatment group as 1, the intercept, .8243, was interpreted as the expected consistency index of the control group in Round 1. Whereas Time had previously been a significant predictor in the unconditional
growth model (Model B2), it became non-significant after adding the Time-by-Group interaction. This interaction, which is the effect under examination for Research Question 1, was significant, explaining about 65% of the variability ($\tau_{11}$) in slopes among individuals. That is, when Group was added to the model, judges’ improvement in their consistency index from one round to the next was better explained than when it was excluded from the model.

The Group variable provided little extra explanatory power to the intercept variance component, $\tau_{00}$. The corresponding pseudo-$R^2$ indicated a reduction in only 2.8% of the unexplained variability, suggesting Group contributed only a small amount to the explanation of the variability of the Round 1 consistency-index values.

Also notable is that the intercept-slope covariance ($\tau_{01}$), which resulted in an intercept-slope correlation of -.008, was not significant. Even though the estimate suggested that judges with lower Round 1 consistency indexes appear to improve in their consistency indexes across rounds at a higher rate than judges with higher Round 1 indexes, this relationship was too weak to be statistically significant.

Compared to the unconditional growth model (Model B2), this model (Model C) fit the data better. The chi-square -2LL deviance ratio test was statistically significant: Even though four degrees of freedom were given up in this model, the decrease in deviance was 34.5 points, which well exceeded the critical chi-square value of 9.49. The low AIC estimate confirmed this conclusion. Because, however, the variance of the slope, .0005, was still statistically significant, I investigated the participant-background covariates to see if these time-invariant predictors removed some of this unexplained variance.
**Participant-background covariate models.** Models D1 – D3 included participant background covariates in the model. I compared each of these with Model C to determine if one or more of these covariates would further explain the trend in improved consistency index from Round 1 to Round 3.

Each of these models consumed one degree of freedom, but added little or no explanatory power to the model. None of the fixed parameters of the time-by-covariate slope was significant and none of the models lowered the unexplained slope variability. Model D2 (with the Assessment experience covariate), for example, was the best of the three covariate models, but added only 1.2% more explanatory power to the variability of the slope (the pseudo-$R^2$ for Model D2’s slope was .665 compared to Model C’s estimate of .653). This slight improvement in fit was not significant, with only a 0.9 improvement, far from the 3.84 critical value needed for chi-square significance. The AIC (-490.5) was also slightly worse than that of Model C (-491.6), providing evidence in favor of the more parsimonious model, Model C. Because I included these covariates to examine whether they confounded the main variable under investigation, the group-by-time interaction, and because they added little explanatory power to the model and resulted in only trivial changes to the model, I excluded them from further analyses and retained Model C as the best working model.

**Results of the first stage of model comparisons.** Model C, which included the logarithmic-trend with the experimental-group predictor and no participant-background covariates, was the best model among the seven presented in Table 4.5. This model predicted a significant Time-by-Group slope effect, which was the parameter under investigation for Research Question 1. The model explained 89% of the within-judge
variance that existed in the unconditional means model, and it explained 65% of the slope variance that existed in the unconditional growth model.

**Covariance structure comparisons.** The fit indexes of the six covariance-structure models are presented in Table 4.6. To identify the most parsimonious and best-fitting covariance structure, I conducted seven comparisons (Table 4.7). Among the five alternative covariance-structure models, TOEPH, TOEP, and AR(1) showed no significant loss in fit when compared with the unstructured model (Comparisons 1, 3, and 4 in Table 4.7); therefore, I compared these models with each other (Comparisons 6 and 7) to identify the best one. The TOEP covariance matrix structure model, which also had the lowest AIC statistic, was the most parsimonious and best-fitting.

Table 4.6

*Fit Indexes of the Logarithmic Fixed-effects Model with Various Covariance Matrix Structures*

<table>
<thead>
<tr>
<th>Covariance structure</th>
<th>Number of covariance parameters</th>
<th>Fit indexes</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>-2LL</td>
</tr>
<tr>
<td>UN</td>
<td>6</td>
<td>-478.2</td>
</tr>
<tr>
<td>TOEPH</td>
<td>5</td>
<td>-478.1</td>
</tr>
<tr>
<td>CSH</td>
<td>4</td>
<td>-455.9</td>
</tr>
<tr>
<td>TOEP</td>
<td>3</td>
<td>-477.6</td>
</tr>
<tr>
<td>AR(1)</td>
<td>2</td>
<td>-473.6</td>
</tr>
<tr>
<td>CS</td>
<td>2</td>
<td>-455.2</td>
</tr>
</tbody>
</table>

*Note.* REML estimation was used for all models. All models were specified as fixed effects with no participant-background covariates. UN = unstructured; TOEPH = heterogeneous Toeplitz; ARH(1) = heterogeneous first-order autoregressive; TOEP = Toeplitz; AR(1) = first-order autoregressive; CS = compound symmetry.
Table 4.7

Comparison of Covariance Structure Models

<table>
<thead>
<tr>
<th>Comparison</th>
<th>Less complex model</th>
<th>More complex model</th>
<th>Δdf</th>
<th>ΔD</th>
<th>Critical chi-square</th>
<th>p</th>
<th>Better model</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>TOEPH</td>
<td>UN</td>
<td>1</td>
<td>0.1</td>
<td>3.84</td>
<td>.752</td>
<td>TOEPH</td>
</tr>
<tr>
<td>2</td>
<td>CSH</td>
<td>UN</td>
<td>2</td>
<td>22.3</td>
<td>5.99</td>
<td>&lt; .001</td>
<td>UN</td>
</tr>
<tr>
<td>3</td>
<td>TOEP</td>
<td>UN</td>
<td>3</td>
<td>0.6</td>
<td>7.81</td>
<td>.896</td>
<td>TOEP</td>
</tr>
<tr>
<td>4</td>
<td>AR(1)</td>
<td>UN</td>
<td>4</td>
<td>4.6</td>
<td>9.49</td>
<td>.331</td>
<td>AR(1)</td>
</tr>
<tr>
<td>5</td>
<td>CS</td>
<td>UN</td>
<td>4</td>
<td>23.0</td>
<td>9.49</td>
<td>&lt; .001</td>
<td>UN</td>
</tr>
<tr>
<td>6</td>
<td>TOEP</td>
<td>TOEPH</td>
<td>2</td>
<td>0.5</td>
<td>5.99</td>
<td>.779</td>
<td>TOEP</td>
</tr>
<tr>
<td>7</td>
<td>AR(1)</td>
<td>TOEP</td>
<td>1</td>
<td>4.0</td>
<td>3.84</td>
<td>.046</td>
<td>TOEP</td>
</tr>
</tbody>
</table>

Note. When ΔD is greater than the critical chi-square value, the more complex model has significantly better fit. ΔD = -1 x [(-2LL of the more complex model) – (-2LL of the less complex model)], which has a chi-square distribution; Δdf = difference in number of parameters between the two models; UN = unstructured; TOEPH = heterogeneous Toeplitz; CSH = heterogeneous compound symmetric; TOEP = Toeplitz; AR(1) = first-order autoregressive; CS = compound symmetric.

Prior to confirming use of the TOEP covariance structure for answering Research Question 1, however, I conducted random-effects model comparisons. Because inclusion of random-effects parameters potentially explains more of the variability, a simpler covariance structure may be more appropriate in random-effects models (Singer, 1998).

Intercepts- and intercepts-and-slopes-as-outcomes models. With the random-effects covariance structure specified as unstructured (in the RANDOM statement in PROC MIXED), none of the random-effects models converged with a positive definite matrix, thereby invalidating parameter estimates. Therefore, following the advice of Fitzmaurice, Laird, and Ware (2011), I specified all random-effects models to have a
factor-analytic random-effects covariance structure, which constrains the random-effects covariance structures to be positive definite.

With the within-subjects covariance matrix structures (in the REPEATED statement in PROC MIXED), I examined three possibilities: TOEP, AR(1), and CS. I selected these because the TOEP structure best fit the fixed-effects model and these three were of equal or greater parsimony. Thus, with the two random-effects models (one with random intercepts and fixed slopes, and the other with random intercepts and slopes) and three within-subjects covariance structures under investigation, there were six possible models.

Only two of the six models converged: One was a random-intercepts model and the other was a random-intercepts-and-slopes model; both had AR(1) covariance structure. Fit statistics indicated worse fit than the fixed-effects model, however, as shown in Table 4.8. That is, neither random-effects model’s deviance was lower than that of the Toeplitz fixed-effects model, which also had the lowest AIC statistic. Thus, the extra random-effects parameters, $\mu_{0j}$ and $\mu_{1j}$, added no explanatory power to the model. The logarithmic fixed-effects model with TOEP covariance structure remained as the best model.

<table>
<thead>
<tr>
<th>Fixed</th>
<th>Random</th>
<th>Repeated-measure covariance structure</th>
<th>Number of parameters</th>
<th>Fit indexes</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>-2LL</td>
</tr>
<tr>
<td>Intercept</td>
<td>Slope</td>
<td>TOEP</td>
<td>9</td>
<td>-477.6</td>
</tr>
<tr>
<td>Slope</td>
<td>Intercept</td>
<td>AR(1)</td>
<td>10</td>
<td>-473.6</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Intercept Slope AR(1)</td>
<td>13</td>
<td>-477.4</td>
</tr>
</tbody>
</table>

*Note.* REML estimation was used for the models. Covariance structure of the random component was specified as factor-analytic. The fixed-effect model is reproduced here to facilitate comparison.
Results of the second stage of model comparisons. Model comparisons showed that the most parsimonious and best fitting model was the fixed-effects logarithmic-trend model with Toeplitz covariance matrix structure (Table 4.9). Before examining the fixed effects of this model, I interpreted the covariance structure and investigated the normality of the residuals.

Table 4.9
Parameter Estimates of the Optimal Model

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Estimate (df)</th>
<th>SE</th>
<th>t</th>
<th>p</th>
<th>LB</th>
<th>UB</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept ($\gamma_{00}$)</td>
<td>.8233 (41.4)</td>
<td>.0096</td>
<td>85.61</td>
<td>&lt; .001</td>
<td>.8038</td>
<td>.8427</td>
</tr>
<tr>
<td>Group ($\gamma_{01}$)</td>
<td>-.0202 (41.4)</td>
<td>.0136</td>
<td>-1.48</td>
<td>.146</td>
<td>-.0476</td>
<td>.0073</td>
</tr>
<tr>
<td>Time ($\gamma_{10}$)</td>
<td>-.0016 (44.2)</td>
<td>.0060</td>
<td>-0.27</td>
<td>.790</td>
<td>-.0138</td>
<td>.0106</td>
</tr>
<tr>
<td>Time*Group ($\gamma_{11}$)</td>
<td>.0659 (44.2)</td>
<td>.0086</td>
<td>7.70</td>
<td>&lt; .001</td>
<td>.0486</td>
<td>.0831</td>
</tr>
<tr>
<td>Covariance structure</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>TOEP(2)</td>
<td>.0015</td>
<td>.0004</td>
<td>4.27</td>
<td>&lt; .001</td>
<td></td>
<td></td>
</tr>
<tr>
<td>TOEP(3)</td>
<td>.0013</td>
<td>.0004</td>
<td>3.54</td>
<td>&lt; .001</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Residual</td>
<td>.0017</td>
<td>.0004</td>
<td>4.70</td>
<td>&lt; .001</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Fit indexes</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>-2LL</td>
<td>-477.6</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>AIC</td>
<td>-471.6</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Note. Parameters were estimated with REML. Degrees of freedom (using the Kenward-Rogers method) of the fixed effects are in parentheses after the parameter estimates. Time was scaled as the natural log of the round.

The covariance parameter test statistics are z values rather than t values.

Covariance matrix structure. The Toeplitz covariance matrix structure placed two constraints on the residual matrix: The variances of the three rounds were set to be equal (homoscedastic) and the correlations between residuals that were one time lag apart.
(between Rounds 1 and 2 and Rounds 2 and 3) were set to be equal. The Toeplitz structure resulted in an estimated covariance matrix (Σ) of

\[
\begin{bmatrix}
.0017 & .0015 \\
.0015 & .0017 \\
.0013 & .0015 & .0017
\end{bmatrix},
\]

which, when translated into a residual correlation matrix, is

\[
\begin{bmatrix}
-.90 & - \\
.74 & .90 & -
\end{bmatrix}.
\]

As would be expected in a repeated-measures design, these correlations decreased as the time lag increased. This reduction was slightly stronger than would have been imposed by the first-order autoregressive structure, which is cited as a commonly used repeated-measures structure (Littell et al., 2006; Singer & Willett, 2003).

**Normality of residuals.** To determine whether the assumption of normality of the residuals held, I examined the skew, kurtosis, Shapiro-Wilk test, and the Q-Q plot. These are presented in Table 4.10 and Figure 4.3. No evidence of deviation from normality was present: The Shapiro-Wilk test was not significant (\(W = 0.99, p = .721\)), the skew and kurtosis confidence intervals contained zero, and the Q-Q plot revealed that residuals did not appear to deviate from normality. With this evidence, I posited that the assumption of normality was met.

**Parameter estimates of the optimal model.** As displayed in Table 4.9, the slope parameter under investigation, \(\gamma_{11}\), was statistically significant and positive, suggesting the null hypothesis for Research Question 1 can be rejected in favor of the first alternative hypothesis, which stated that judges receiving feedback will show greater gains in consistency across rounds. The two main effects (of Time and Group) were not found to be significant.
Table 4.10

*Normality of the Residuals of the Optimal Model*

<table>
<thead>
<tr>
<th></th>
<th>Kurtosis</th>
<th></th>
<th></th>
<th>Skewness</th>
<th></th>
<th></th>
<th>Shapiro-Wilk test</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$M$</td>
<td>$SD$</td>
<td>Estimate</td>
<td>$LB$</td>
<td>$UB$</td>
<td>Estimate</td>
<td>$LB$</td>
</tr>
<tr>
<td>0.00</td>
<td>0.04</td>
<td>-1.27</td>
<td>-1.06</td>
<td>0.81</td>
<td>.089</td>
<td>-0.38</td>
<td>0.56</td>
</tr>
</tbody>
</table>

*Note.* Confidence intervals of the kurtosis and skewness estimates were calculated by multiplying 1.98 (the $t$-value at the 95% confidence limit with 108 data points) by the standard error (SE) of each estimate, $SE_{kurt} = \sqrt{24/108}$ and $SE_{skew} = \sqrt{6/108}$. LB = lower bound 95% confidence limit; UB = upper bound 95% confidence limit.

*Figure 4.3.* Quantile-quantile plot of the residuals of the optimal model. Mu = $M$; sigma = $SD$. 

![Q-Q Plot of the Residuals](image-url)
With the parameter estimates in the modeling equation, the model takes the following form:

\[ C_{jr} = 0.8233 - 0.0202(\text{Group})_j - 0.0016(\ln\text{Round})_{jr} + 0.0659(\ln\text{Round}*\text{Group})_{jr} + r_{jr}, \]

where \( C_{jr} \) is the consistency index and \( r_{jr} \) is the covariance matrix structure reported in Table 4.9. With the control group, which is coded as \( \text{group} = 0 \), the terms with the group variable reduce to zero. Thus, the intercept, 0.8233, is an estimate of the control group’s consistency index during Round 1 (when \( \ln\text{Round} = 0 \)). The slope (of the interaction) is interpreted as the following: From Round \( r - 1 \) to Round \( r \), judges receiving the feedback are expected to increase in intrajudge consistency by \( 0.0659 \times [\ln(r) - \ln(r - 1)] \) more points than judges not receiving the feedback. For example, from Round 1 to Round 2, the expected effect of receiving over not receiving feedback is \( 0.0659 \times [\ln(2) - \ln(2 - 1)] \) = \( 0.0659 \times [0.69 - 0] = 0.0457 \) points; from Round 2 to Round 3, the expected effect is \( 0.0659 \times [1.10 - 0.69] = 0.0267 \) points. Thus, the expected trend is that if intrajudge-consistency feedback is provided between each round, its positive effect will diminish asymptotically as rounds progress. The model with the data under investigation is illustrated in Figure 4.4.

**Generalizability-theory Results for Examining Interjudge Consistency**

Research Questions 2 and 3 required an examination of the variability of the Angoff ratings. To facilitate interpretation of the generalizability-theory results when they are translated into raw score points, I calculated the descriptive statistics of the cutscores. Before running the descriptive statistics analysis, I calculated each judge’s cutscore as the sum of the Angoff ratings across the set of items.
Descriptive statistics of the cutscores. Table 4.11 displays the descriptive statistics of the cutscores for each of the two experimental groups in each of the three rounds. Neither group changed their recommended cutscore across the three rounds of rating by more than 0.40 of a point on the 0 to 50 raw-score scale. The two group’s recommended cutscores were similar to each other, differing by no more than 1.30 points, far lower than either group’s standard deviations. The treatment group’s standard deviations decreased slightly from an initial 6.33 in Round 1 to 5.86 in Round 3, indicating judges’ cutscore variability decreased by about half a raw-score point, which suggested judges in this group slightly improved in their degree of agreement in where to place the cutscore. This pattern was not observed with the control group, whose standard deviations increased slightly from 7.06 in Round 1 to 7.42, resulting in about a third of a point increase in divergence.

Figure 4.4. Plot of the fixed-effects logarithmic model with Toeplitz covariance structure.
Table 4.11

*Descriptive Statistics of the Cutscores per Group per Round*

<table>
<thead>
<tr>
<th>Group</th>
<th>Round</th>
<th>M</th>
<th>SD</th>
<th>Low</th>
<th>Median</th>
<th>High</th>
</tr>
</thead>
<tbody>
<tr>
<td>Treatment</td>
<td>1</td>
<td>33.42</td>
<td>6.33</td>
<td>19.90</td>
<td>34.33</td>
<td>45.71</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>33.22</td>
<td>5.91</td>
<td>20.00</td>
<td>34.79</td>
<td>45.89</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>33.02</td>
<td>5.86</td>
<td>19.65</td>
<td>34.79</td>
<td>45.91</td>
</tr>
<tr>
<td>Control</td>
<td>1</td>
<td>34.41</td>
<td>7.06</td>
<td>14.87</td>
<td>36.10</td>
<td>42.78</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>34.32</td>
<td>7.28</td>
<td>13.52</td>
<td>35.86</td>
<td>42.88</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>34.29</td>
<td>7.42</td>
<td>12.49</td>
<td>35.20</td>
<td>43.11</td>
</tr>
</tbody>
</table>

*Note.* Cutscores are on the raw scale, with a possible range of 0 to 50. The descriptive statistics of the mean Angoff ratings across items are calculable by dividing these estimates by 50.

**G-study results.** To answer Research Question 2, I conducted a G study to investigate the degree to which judges within each experimental group converged in their Angoff ratings. Results here show the percentage of variance of each component, the dependability index, and the estimated standard errors of the cutscore when the item component is treated as random, $\hat{\sigma}_X$, and fixed, $\hat{\sigma}_{X|I}$.

**Variance components.** The variance components of the treatment group, presented in Table 4.12, revealed that the estimated item-by-judge variance decreased from about 36% to 23% of the total variance from Round 1 to Round 3, indicating a reduction in unexplained or idiosyncratic variability. Variability in the judges’ mean Angoff ratings, however, decreased only slightly, explaining about 20% of the variability in Round 1 and 19% in Round 3. These two variance estimates suggested that the treatment-group judges improved in their degree of agreement at the item level but that this was not reflected at the cutscore level (mean-across-item level). The item variance had the opposite pattern to that of the error variance, increasing from about 44% in Round 1 to about 58% in
Round 3, indicating that Angoff ratings became more spread out as judges moved from Round 1 to Round 3. This change suggested that judges’ appeared to conceptualize a broader range of item difficulty.

Table 4.12
*Treatment Group’s Estimated Variance Components in Each of the Three Rounds*

<table>
<thead>
<tr>
<th>Round</th>
<th>Source of variance</th>
<th>Mean square</th>
<th>Estimated variance component</th>
<th>Percent of variance within the round</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Item ($i$)</td>
<td>0.6519</td>
<td>0.0347</td>
<td>44.48</td>
</tr>
<tr>
<td></td>
<td>Judge ($j$)</td>
<td>0.8021</td>
<td>0.0155</td>
<td>19.87</td>
</tr>
<tr>
<td></td>
<td>$ij,e$</td>
<td>0.0278</td>
<td>0.0278</td>
<td>35.64</td>
</tr>
<tr>
<td>2</td>
<td>Item ($i$)</td>
<td>0.6895</td>
<td>0.0373</td>
<td>53.86</td>
</tr>
<tr>
<td></td>
<td>Judge ($j$)</td>
<td>0.6996</td>
<td>0.0136</td>
<td>19.69</td>
</tr>
<tr>
<td></td>
<td>$ij,e$</td>
<td>0.0183</td>
<td>0.0183</td>
<td>26.45</td>
</tr>
<tr>
<td>3</td>
<td>Item ($i$)</td>
<td>0.7463</td>
<td>0.0406</td>
<td>58.12</td>
</tr>
<tr>
<td></td>
<td>Judge ($j$)</td>
<td>0.6871</td>
<td>0.0134</td>
<td>19.24</td>
</tr>
<tr>
<td></td>
<td>$ij,e$</td>
<td>0.0158</td>
<td>0.0158</td>
<td>22.64</td>
</tr>
</tbody>
</table>

*Note.* Degrees of freedom of the items was 49, of the judges was 17, and of the interaction was 833. The outcome variable was the Angoff rating, in probability units. The item component was the object of measurement.

For the control group the estimated item-by-judge variance changed from 32% to 31%, suggesting negligible reduction in unexplained variance (Table 4.13). Variability in the judge facet increased moderately from about 30% to 35%. This pattern differed from that of the treatment group in that control-group judges showed virtually no improved agreement in their item-level judgments and further diverged in their cutscore estimates as the rounds progressed. Variability in the items decreased from 38% to about 34%, suggesting judges used less of the range in item difficulty.
Table 4.13

Control Group’s Estimated Variance Components in Each of the Three Rounds

<table>
<thead>
<tr>
<th>Round</th>
<th>Source of variance</th>
<th>Mean square</th>
<th>Estimated variance component</th>
<th>Percent of variance within the round</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Item (i)</td>
<td>0.4609</td>
<td>0.0244</td>
<td>37.56</td>
</tr>
<tr>
<td></td>
<td>Judge (j)</td>
<td>0.9957</td>
<td>0.0195</td>
<td>29.97</td>
</tr>
<tr>
<td></td>
<td>ij,e</td>
<td>0.0211</td>
<td>0.0211</td>
<td>32.47</td>
</tr>
<tr>
<td></td>
<td>Item (i)</td>
<td>0.4247</td>
<td>0.0225</td>
<td>36.00</td>
</tr>
<tr>
<td>2</td>
<td>Judge (j)</td>
<td>1.0586</td>
<td>0.0208</td>
<td>33.24</td>
</tr>
<tr>
<td></td>
<td>ij,e</td>
<td>0.0192</td>
<td>0.0192</td>
<td>30.76</td>
</tr>
<tr>
<td></td>
<td>Item (i)</td>
<td>0.4051</td>
<td>0.0214</td>
<td>34.45</td>
</tr>
<tr>
<td>3</td>
<td>Judge (j)</td>
<td>1.1000</td>
<td>0.0216</td>
<td>34.74</td>
</tr>
<tr>
<td></td>
<td>ij,e</td>
<td>0.0192</td>
<td>0.0192</td>
<td>30.82</td>
</tr>
</tbody>
</table>

Note. Degrees of freedom of the items was 49, of the judges was 17, and of the interaction was 833. The outcome variable was the Angoff rating, in probability units. The item component was the object of measurement. The $ij,e$ variance is the sum of the interaction and error variance components.

**Dependability indexes.** The dependability indexes, reported in Table 4.14, provided measures of the generalizability of the procedure for making absolute decisions about each item’s conditional difficulty. That is, the index provided an estimate of the consistency of a population of judges’ item-level Angoff ratings, regardless of the relative difficulty items in the set. These results were consistent with the results in the variance-component analysis: the dependability indexes appeared to improve in the treatment group but worsen in the control group.

**Standard errors of the mean cutscore.** Measures of the standard error of the cutscore, in Table 4.14, provided estimates of the convergence in judges’ ratings in terms of precision. When both items and judges were assumed to be random, neither group showed improvement in their cutscore precision: the treatment group’s standard error in raw-score points was estimated as 1.99 in Round 1 and 1.98 in Round 3; that of the
control group increased (worsened) slightly from 2.00 to 2.03. When judges are assumed to be random and items fixed, the control group’s precision slightly worsened by about 0.09 of a raw-score point, whereas the treatment group’s slightly improved by 0.11 of a raw-score point.

Table 4.14

*Dependability Indexes and Estimates of the Standard Error of the Cutscore*

| Experimental group | Round | Dependability index | $\hat{\sigma}_X$ (raw) | $\hat{\sigma}_{X|I}$ (raw) |
|--------------------|-------|---------------------|-------------------------|---------------------------|
| Treatment          | 1     | 0.935               | 0.0398 (1.99)           | 0.0299 (1.49)             |
|                    | 2     | 0.955               | 0.0390 (1.95)           | 0.0279 (1.39)             |
|                    | 3     | 0.962               | 0.0397 (1.98)           | 0.0276 (1.38)             |
| Control            | 1     | 0.915               | 0.0399 (2.00)           | 0.0333 (1.66)             |
|                    | 2     | 0.910               | 0.0403 (2.02)           | 0.0343 (1.71)             |
|                    | 3     | 0.904               | 0.0406 (2.03)           | 0.0350 (1.75)             |

*Note.* Estimates of the standard errors in raw-score units are in parentheses. $\hat{\sigma}_X$ = standard error of the cutscore when items and judges are treated as random samples of the domain and population. $\hat{\sigma}_{X|I}$ = standard error of the cutscore when judges are considered random but items are fixed.

**D-study results.** The D-study results based on Hurtz and Hertz’s (1999) criterion (a standard error, $\hat{\sigma}_X$, that is less than 2.5 raw-score units) provided a translation of the earlier G-study standard-error results (when items and judges are both assumed to be random) into a number-of-judges metric. The results of Round 1, presented in Table 4.15, suggested that about nine treatment- or 10 control-group judges would be needed if a single round of Angoff rating were conducted. The results of Round 3, presented in Table 4.16, indicated that the treatment group’s precision was the same as that in Round 1 (nine judges would be needed), but that the control group’s precision worsened by about one judge (11 judges would be needed). This very small change was consistent with the
G-study results (in that neither group changed much in their cutscore precision when both items and judges were assumed to be random components). Examination of the diminishing returns with the Round 3 data, as illustrated in Figure 4.5, revealed that after about 10 or 11 judges, little improvement in the standard error of the cutscore, for both experimental groups, appears to be achieved.

The D-study results based on the .80 dependability-index criterion provided a translation of the earlier G-study dependability-index results into a number-of-judges metric. With the treatment group, it was estimated that in Round 1 at least five judges would be needed. In Round 3, three treatment-group judges would be needed. With the control group, it was estimated that seven judges would be needed in Round 1 and eight judges in Round 3. Thus, future scenarios employing the .80 dependability-index criterion will likely achieve adequate precision with around three to eight judges, depending on whether intrajudge-consistency feedback is provided. With judges receiving the feedback, the precision was estimated to improve by an amount that is equivalent to about two judges; in groups with judges not receiving feedback, the precision might be expected to worsen by about one judge. Figure 4.6 illustrates the Round 3 data in terms of the diminishing returns of adding more judges. It is safe to say that after about 10 or 11 treatment-group judges, little improvement in the dependability indexes appears to be achieved by hiring more participants.
Table 4.15

*D Study of the Round 1 Ratings*

<table>
<thead>
<tr>
<th>N judges</th>
<th>Treatment group</th>
<th>Control group</th>
<th>Dependability</th>
<th>Dependability</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$\hat{\sigma}_1^2$</td>
<td>$\hat{\sigma}_{i_1}^2$</td>
<td>$\hat{\sigma}_J^2$ (raw)</td>
<td>$\hat{\sigma}_1^2$</td>
</tr>
<tr>
<td>20</td>
<td>.0008</td>
<td>.0014</td>
<td>.0387 (1.93)</td>
<td>.941</td>
</tr>
<tr>
<td>19</td>
<td>.0008</td>
<td>.0015</td>
<td>.0392 (1.96)</td>
<td>.938</td>
</tr>
<tr>
<td>18</td>
<td><strong>.0009</strong></td>
<td><strong>.0015</strong></td>
<td><strong>.0398 (1.99)</strong></td>
<td><strong>.935</strong></td>
</tr>
<tr>
<td>17</td>
<td>.0009</td>
<td>.0016</td>
<td>.0405 (2.02)</td>
<td>.932</td>
</tr>
<tr>
<td>16</td>
<td>.0010</td>
<td>.0017</td>
<td>.0412 (2.06)</td>
<td>.928</td>
</tr>
<tr>
<td>15</td>
<td>.0010</td>
<td>.0019</td>
<td>.0420 (2.10)</td>
<td>.923</td>
</tr>
<tr>
<td>14</td>
<td>.0011</td>
<td>.0020</td>
<td>.0429 (2.14)</td>
<td>.918</td>
</tr>
<tr>
<td>13</td>
<td>.0012</td>
<td>.0021</td>
<td>.0439 (2.20)</td>
<td>.912</td>
</tr>
<tr>
<td>12</td>
<td>.0013</td>
<td>.0023</td>
<td>.0451 (2.25)</td>
<td>.906</td>
</tr>
<tr>
<td>11</td>
<td>.0014</td>
<td>.0025</td>
<td>.0464 (2.32)</td>
<td>.898</td>
</tr>
<tr>
<td>10</td>
<td>.0015</td>
<td>.0028</td>
<td>.0479 (2.40)</td>
<td>.889</td>
</tr>
<tr>
<td>9</td>
<td>.0017</td>
<td>.0031</td>
<td>.0498 (2.49)$^\dagger$</td>
<td>.878</td>
</tr>
<tr>
<td>8</td>
<td>.0019</td>
<td>.0035</td>
<td>.0519 (2.60)</td>
<td>.865</td>
</tr>
<tr>
<td>7</td>
<td>.0022</td>
<td>.0040</td>
<td>.0546 (2.73)</td>
<td>.849</td>
</tr>
<tr>
<td>6</td>
<td>.0026</td>
<td>.0046</td>
<td>.0580 (2.90)</td>
<td>.828</td>
</tr>
<tr>
<td>5</td>
<td>.0031</td>
<td>.0056</td>
<td>.0625 (3.12)</td>
<td>.800$^\ddagger$</td>
</tr>
<tr>
<td>4</td>
<td>.0039</td>
<td>.0069</td>
<td>.0686 (3.43)</td>
<td>.762</td>
</tr>
<tr>
<td>3</td>
<td>.0052</td>
<td>.0093</td>
<td>.0777 (3.89)</td>
<td>.706</td>
</tr>
<tr>
<td>2</td>
<td>.0077</td>
<td>.0139</td>
<td>.0934 (4.67)</td>
<td>.616</td>
</tr>
<tr>
<td>1</td>
<td>.0155</td>
<td>.0278</td>
<td>.1294 (6.47)</td>
<td>.445</td>
</tr>
</tbody>
</table>

*Note.* The boldfaced row indicates the observed estimates. The estimated item variance component, $\hat{\sigma}_i^2$, was .0347 for all treatment-group scenarios and .0244 for all control-group scenarios; $\hat{\sigma}_J^2$ = estimated variance averaged across the mean of judges; $\hat{\sigma}_{i_1}^2$ = estimated variance components of the interaction; $\hat{\sigma}_X^2$ = Estimated standard error of the cutscore, with random items and judges.

$^\dagger$Met Hurtz and Hertz’s (1999) criterion.

$^\ddagger$Met the .80 dependability-index criterion.
Table 4.16

*D Study of the Round 3 Ratings*

<table>
<thead>
<tr>
<th>N judges</th>
<th>Treatment group</th>
<th>Control group</th>
<th>Dependability index</th>
<th>Dependability index</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$\hat{\sigma}_I^2$</td>
<td>$\hat{\sigma}_{II,e}^2$</td>
<td>$\hat{\sigma}_{X}^2$ (raw)</td>
<td>$\hat{\sigma}_I^2$</td>
</tr>
<tr>
<td>20</td>
<td>.0007</td>
<td>.0008</td>
<td>.0387 (1.94)</td>
<td>.965</td>
</tr>
<tr>
<td>19</td>
<td>.0007</td>
<td>.0008</td>
<td>.0392 (1.96)</td>
<td>.965</td>
</tr>
<tr>
<td>18</td>
<td><strong>.0007</strong></td>
<td><strong>.0009</strong></td>
<td><strong>.0397 (1.98)</strong></td>
<td><strong>.962</strong></td>
</tr>
<tr>
<td>17</td>
<td>.0008</td>
<td>.0009</td>
<td>.0402 (2.01)</td>
<td>.959</td>
</tr>
<tr>
<td>16</td>
<td>.0008</td>
<td>.0010</td>
<td>.0409 (2.04)</td>
<td>.957</td>
</tr>
<tr>
<td>15</td>
<td>.0009</td>
<td>.0011</td>
<td>.0416 (2.08)</td>
<td>.954</td>
</tr>
<tr>
<td>14</td>
<td>.0010</td>
<td>.0011</td>
<td>.0423 (2.12)</td>
<td>.951</td>
</tr>
<tr>
<td>13</td>
<td>.0010</td>
<td>.0012</td>
<td>.0432 (2.16)</td>
<td>.947</td>
</tr>
<tr>
<td>12</td>
<td>.0011</td>
<td>.0013</td>
<td>.0442 (2.21)</td>
<td>.943</td>
</tr>
<tr>
<td>11</td>
<td>.0012</td>
<td>.0014</td>
<td>.0454 (2.27)</td>
<td>.939</td>
</tr>
<tr>
<td>10</td>
<td>.0013</td>
<td>.0016</td>
<td>.0468 (2.34)</td>
<td>.933</td>
</tr>
<tr>
<td>9</td>
<td>.0015</td>
<td>.0018</td>
<td>.0484 (2.42)(†)</td>
<td>.926</td>
</tr>
<tr>
<td>8</td>
<td>.0017</td>
<td>.0020</td>
<td>.0503 (2.51)</td>
<td>.917</td>
</tr>
<tr>
<td>7</td>
<td>.0019</td>
<td>.0023</td>
<td>.0527 (2.63)</td>
<td>.907</td>
</tr>
<tr>
<td>6</td>
<td>.0022</td>
<td>.0026</td>
<td>.0557 (2.78)</td>
<td>.893</td>
</tr>
<tr>
<td>5</td>
<td>.0027</td>
<td>.0032</td>
<td>.0597 (2.98)</td>
<td>.874</td>
</tr>
<tr>
<td>4</td>
<td>.0034</td>
<td>.0040</td>
<td>.0652 (3.26)</td>
<td>.847</td>
</tr>
<tr>
<td>3</td>
<td>.0045</td>
<td>.0053</td>
<td>.0734 (3.67)</td>
<td>.806(‡)</td>
</tr>
<tr>
<td>2</td>
<td>.0067</td>
<td>.0079</td>
<td>.0877 (4.38)</td>
<td>.735</td>
</tr>
<tr>
<td>1</td>
<td>.0134</td>
<td>.0158</td>
<td>.1206 (6.03)</td>
<td>.581</td>
</tr>
</tbody>
</table>

Note. The boldfaced row indicates the observed estimates. The estimated item variance component, $\hat{\sigma}_I^2$, was .0406 for all treatment-group scenarios and .0214 for all control-group scenarios; $\hat{\sigma}_{II,e}^2$ = estimated variance averaged across the mean of judges; $\hat{\sigma}_{II,e}^2$ = estimated variance components of the interaction; $\hat{\sigma}_{X}^2$ = Estimated standard error of the cutscore, with random items and judges.

\(†\)Met Hurtz and Hertz’s (1999) criterion.

\(‡\)Met the .80 dependability-index criterion.
Figure 4.5. D-study plot of the Round 3 standard error of the cutscore estimates with varying numbers of judges. Item and judge components are assumed to be random.

Figure 4.6. D-study plot of the Round 3 dependability-index estimates with varying numbers of judges.
CHAPTER 5
DISCUSSION

Determining where to place a cutscore on a test requires human judgment. The Angoff method elicits this human judgment at the item level and yields an estimate of a judge’s recommended cutscore from the sum of these judgments across items. The recommended cutscore of a panel of judges is the mean (sometimes the median) across judges. One concern, argued by some critics of the Angoff method as grounds for abandoning the approach altogether (Shepard, 1995; Shepard et al., 1993), is that judges may have difficulty maintaining a stable conceptualization of the BPE while reviewing multiple test items. Another concern, relevant to consumers of the cutscores, is the degree of confidence that can be placed in a panel’s recommended cutscore when there is variability among judges in their Angoff ratings and cutscores (Hambleton & Pitoniak, 2006; Hurtz & Hertz, 1999; Plake, 2008). These two concerns can be translated into two kinds of error: intrajudge and interjudge inconsistency, which can be measured using van der Linden’s (1982) intrajudge-consistency index and estimates of variability in G theory (Brennan & Lockwood, 1980), respectively. The purpose of this dissertation is to examine the effect of intrajudge-consistency feedback on these two kinds of error.

Item-level intrajudge-consistency feedback is often provided between rounds for the purpose of improving panelists’ judgments and ensuring their ratings reflect their true conceptualizations of the performance standard (Reckase, 2001; Reckase & Chen, 2012). Provision of this feedback is in accord with the Standards’ (AERA, APA, & NCME, 1999) recommendation that standard-setting procedures “…be designed so that judges can bring their knowledge and experience to bear in a reasonable way [and that these...
procedures] result in reasonable, defensible standards that accurately reflect the judges’ values and intentions” (p. 60). In this way, including intrajudge-consistency feedback in a standard-setting procedure strengthens the procedural aspect of the validity argument. Likewise, evidence showing the effectiveness of the feedback can strengthen internal validity.

**Intrajudge Consistency**

My first research question called for an investigation into whether item-level intrajudge-consistency feedback has an effect on panelists’ intrajudge consistency. To address this, I compared panelists receiving item-level intrajudge-consistency feedback with those participating in an alternative, non-feedback, activity. Specifically, I compared these two groups’ round-to-round changes in consistency indexes by testing the significance of the Time-by-Group ($\gamma_{11}$) parameter under a multilevel-model-for-change framework. The results showed strong support for my first hypothesis, that the feedback yields significant improvement: For the optimal (best fitting and most parsimonious) model, $t(44.2) = 7.70, p < .001$. This model predicts that judges receiving the feedback will improve more than the judges not receiving the feedback and that this improved gain is estimated to be .0457 consistency-index points when going from Round 1 to Round 2 and about .0267 points when going from Round 2 to Round 3.

**Strength of the effect.** One way to put this predicted treatment-versus-control change into perspective is to compare the improvement to the standard deviations of the observed consistency indexes. The predicted Round 1 to Round 2 improvement (.0457) is nearly equivalent in magnitude to the control group’s Round 1 standard deviation (.0463), which was the highest among the six sets of data (the three rounds for each of the two
experimental groups, reported in Table 4.1). This means that the first between-round feedback has the effect of improving consistency by around one standard-deviation unit. The second between-round feedback is estimated to further improve consistency, by more than half of the magnitude of the first improvement.

Another way to examine the strength of this improvement is to compare it to consistency indexes reported in previous studies (presented in Table 2.1; Chang, van der Linden, & Vos, 2004; Friedman & Ho, 1990; Hurtz & Jones, 2009; van der Linden, 1982). The three studies using data based on actual judges (Chang et al., 2004; Friedman & Ho, 1990; van der Linden, 1982) reported standard deviations that ranged from .018 to .041—estimates that are lower than the predicted .0457 Round 1 to Round 2 gain in this study. In the two studies (Friedman & Ho, 1990; Hurtz & Jones, 2009) that reported a second round of rating (none reported a third), the intrajudge-consistency indexes increased by .030 from Round 1 to Round 2. Thus, my study predicts improvement beyond previous studies’ gains by about .015 consistency-index units.

The between-round activities in these two studies (Friedman & Ho, 1990; Hurtz & Hertz, 1999), however, did not include intrajudge-consistency feedback. Hurtz and Jones’ (2009) Round 2 data were simulated (from their Round 1 simulation) to approximate the effect that they speculated would be due to usual between-round feedback and discussion activities used in Angoff methods. Friedman and Ho’s (1990) between-round feedback involved judges’ discussing and reviewing each other’s item-level ratings. Thus, the consistency-index gains in these two studies were not due to intrajudge-consistency feedback. To the best of my knowledge, my study is the first to examine this effect on this dependent variable.
Still another way to examine the strength of this improvement is to consider the proportional reduction in residual variance (pseudo-$R^2_{τ11}$) of the judges’ estimated consistency indexes across rounds after adding Group to the unconditional growth model. The pseudo-$R^2_{τ11}$ was calculable only in the first stage of the model comparisons, before the within-judge covariance structure was specified; therefore, the estimate should be taken with caution. This drawback aside, the strength of the effect was substantial, at pseudo-$R^2_{τ11} = .653$, which suggests that the feedback explained approximately 65% of the variability in judges’ round-to-round changes in intrajudge consistency. Because the fixed-effects estimates and standard errors of Model C (in the first stage of model comparisons) were similar to those estimated in the optimal model, it is reasonable to assume that the optimal model explained a similar amount of variability.

Given that the feedback provided to the treatment-group judges was non-numerical, this effect cannot be attributed to what Clauser, Mee, et al. (2009) characterized as “mechanical” reliance on the provided conditional $p$-values. There were no recommended $p$-values or Angoff ratings in the feedback. Instead, the judges were instructed to more closely examine 12 items that were flagged as least consistent. They were told that their rating on the item was somewhat or very inconsistent, depending on the degree of their inconsistency, and they were told that likely more or fewer of the BPEs would have gotten the item correct, depending on the direction of their inconsistency. Thus, whereas Clauser, Mee, et al.’s caution might apply to Reckase Charts, which provide numerical intrajudge-consistency feedback, it did not apply to the current study.

*Generalizing to numerical feedback procedures.* In previous inquiries into the use of non-numerical intrajudge-consistency feedback, described in Loomis’s (2000) report of
the NAEP standard-setting development, judges provided self-reports that suggested they either ignored or misunderstood the feedback. Data in the current study suggest otherwise. If judges in my study had not received the feedback, they likely would not have improved, as suggested by the lack of gains by the control group. The agency managing the NAEP standard-setting procedures in the nineties (ACT) replaced non-numerical feedback with Reckase Charts based on these self-report data (Loomis, 2000).

If Reckase Charts indeed provide more understandable and usable information than non-numerical feedback, as suggested by Loomis (2000), and numerical data have a stronger influence on judges’ ratings, as found by Clauser, Mee, et al. (2009), it is likely that numerical feedback would yield an even stronger gain than that observed in my study. Further support of this reasoning is that judges in my study reviewed feedback on only 12 (24%) of the 50 items (between each round), whereas the numerical intrajudge-consistency feedback presented in Reckase Charts typically shows the judge’s location on all items. Thus, it is not unreasonable to assume that the results of my study provide a low estimate of the effect we would expect with numerical feedback. Depending on stakeholders’ philosophical position on what the nature of the feedback provided to judges should be (as discussed in Reckase & Chen, 2012), numerical feedback data may be acceptable. The strong effect observed in my study provides empirical support for either numerical or non-numerical intrajudge-consistency feedback, satisfying either end of this philosophy spectrum.
Interjudge Consistency

My second research question called for an investigation into whether judges receiving intrajudge-consistency feedback show improved interjudge consistency, operationalized as reduced variability among the judges in their cutscores and item-level ratings. The findings provide support for using intrajudge-consistency feedback to improve this aspect of a standard-setting procedure’s credibility.

Changes in variance components. The G-study results in this study are consistent with those discussed in studies investigating the effect of conditional $p$-value feedback (which is similar to numerical intrajudge consistency feedback) on variance components. Similar to the findings in Clauser et al. (2002), Clauser, Harik, et al. (2009), and Dillon and Walsh (2000), the results of my study suggest that the error variance, $\hat{\sigma}_{ij,e}^2$, decreases somewhat (explaining 36% in Round 1, but 23% in Round 3) after feedback is provided. That is, judges likely become less idiosyncratic in their item-level judgments after reviewing intrajudge-consistency feedback. Additionally, the pattern of the control-group judges’ error variance suggests that judges not provided with feedback may show only negligible improvement (with $\hat{\sigma}_{ij,e}^2$ explaining 32% in Round 1, and 31% in Round 3). Thus, it is reasonable to infer that item-level intrajudge-consistency feedback functions in a way that is similar to conditional $p$-value feedback in that it reduces the variability attributed to idiosyncratic judgments.

The changes in amount of variance explained by the items, $\hat{\sigma}_i^2$, suggested that the feedback improved one aspect of the internal validity of the procedure. For the treatment group, this variance component increased, suggesting that judges use a broader range of item difficulty after receiving feedback. This can be interpreted to mean that intrajudge-
consistency feedback contributes to the validity of the procedure by abating the concern that Angoff judges might overestimate success on difficult items and underestimate success on easy items, an issue raised by critics (Shepard, 1995; Shepard et al., 1993) of the Angoff method.

The variance among judges, $\hat{\sigma}_j^2$, in the treatment group decreased only a small amount from Round 1 to Round 3, which may be interpreted to mean that feedback probably has little effect on improving judges’ consensus in their conceptualizations of the performance standard. However, this variance component increased in the control group, suggesting that when judges are left to their own devices, they may become more divergent in their conceptualizations of the BPE’s ability. Thus, not providing feedback seems to result in worsening of judge consensus.

**Dependability indexes.** Whereas the control group worsened in dependability-index estimates from .915 to .904, the treatment group improved from .935 to .962. This suggests that intrajudge-consistency feedback improves the dependability of the Angoff procedure.

This improved dependability was very similar to that derived from variance-component estimates reported in Clauser, Harik, et al. (2009): In their two studies employing conditional performance-data feedback, the dependability improved from .934 to .961 and from .928 to .969. Thus, the non-numerical feedback employed in my study has about the same effect as the numerical conditional $p$-values provided in Clauser, Harik, et al.

Even before feedback or other between-round activities were provided, the dependability indexes were high, exceeding .90 for both experimental groups. Hurtz and
Hertz (1999) characterized a dependability index of .87 as “quite high”, as it was the highest they observed among the eight Angoff procedures they examined. Thus, the indexes in my study are comparatively substantial. Because high indexes can occur when there is a large range in the items’ empirical difficulty (a point brought up by Raymond & Reid, 2001), however, comparisons across studies should be interpreted with caution. Nonetheless, the high dependability indexes observed in my study suggest that judges are sensitive to variability in the items on the VST and that their ratings are highly dependable (regardless of whether they got the feedback).

**Standard errors of the cutscore with random items.** For evaluating the precision of the cutscore when generalizing the result to other Angoff procedures in which a similar sample of judges are asked to rate a similar sample of items, the changes in $\sigma_X$ (the standard error of the cutscore when the item and judge components are treated as random) from Round 1 to Round 3 are informative. The data in my study suggest that for both experimental groups, these changes were negligible. Thus, intrajudge-consistency feedback (and the iterative rating process, for that matter) appears to contribute little to the precision of the cutscore when the cutscore is assumed to generalize to future standard-setting conditions in which similar samples of judges rate similar samples of vocabulary items.

Decisions about the use of the final-round recommended cutscore (which, on the 50-point scale, was 34.29 for the control group and 33.02 for the treatment group) can be based on these standard error estimates: The Round 3 $\hat{\sigma}_X$ (which, in raw-score units, was 2.03 for the control and 1.98 for the treatment group) can be translated into 95% confidence intervals for making decisions based on the cutscore estimates with future
versions of the VST. For the treatment group, the interval in raw-score points was from 28.83 to 37.21; that of the control group was similar, with an interval of 30.00 to 38.57. Consumers (such as school personnel or teachers) should take this interval into account when using other versions of the VST (that are not equated or demonstrated to be parallel to the item set in this study) to identify students who have enough vocabulary to be considered prepared for the demands of college reading.

Ultimately, determining how to use the standard error of the cutscore is a policy decision (Cizek & Bunch, 2007). The broad band of $\sigma_X$ uncertainty may be acceptable to consumers of this cutscore recommendation because its use is for low stakes. Additionally, because the cutscore is intended to be used for identifying students who need more vocabulary instruction, consumers might place the operational cutscore at the high end of this band, perhaps at one $\sigma_X$ above the recommended cutscore. This would yield a functioning cutscore of 35 if using the treatment group’s cutscore on the 50-item test (which would be a score of 98 on the full 140-item version of the VST). For more cautious decision-making with this intended use, consumers might place the operational cutscore at 37 points (equivalent to 104 points on the full 140-item VST), which is near the top of the 95% confidence band. Additionally, the precision (such as the standard error of measurement) of the particular version of the VST that is used should be taken into account when making classification decisions (Cizek & Bunch, 2007; Geisinger, 1991; Jaeger, 1991).

When examining the $\sigma_X$ in probability units, it is possible to compare this study’s precision with that of other studies (e.g., Brennan & Lockwood, 1980; Cross et al., 1984; Fehrmann et al., 1991; Hurtz & Hertz, 1999) that assume items are a random component.
On this scale, my study’s estimates range from .0390 (in Round 2 of the treatment group) to .0406 (in Round 3 of the control group). The mean pre-feedback $\hat{\sigma}_X$ across three studies (Brennan & Lockwood, 1980; Cross et al., 1984; Fehrmann et al., 1991) was .0373, which is slightly lower than the estimates in my study. The mean post-feedback $\hat{\sigma}_X$ across two studies (Cross et al., 1984; Hurtz & Hertz, 1999) was .0433, which is slightly higher than those in my study. Thus, the estimated standard errors of the cutscore in my study seem to fall reasonably close to those that already exist in the literature.

**Standard errors of the cutscore with fixed items.** For evaluating the precision of the cutscore when generalizing the result to other Angoff procedures in which a similar sample of judges are asked to rate the same set of items, the changes in $\hat{\sigma}_{X|I}$ (the standard error of the cutscore when the judge component is treated as random but the item component is treated as fixed) from Round 1 to Round 3 are informative. The data in my study suggest there were changes over the rounds and that the experimental groups did differ from each other. The $\hat{\sigma}_{X|I}$ improved for the treatment group (from 1.49 to 1.38 raw-score units), but worsened slightly for the control group (from 1.66 to 1.75). Even though these changes were only about a tenth of a raw-score point in magnitude, the data do suggest there was some improvement in consensus among judges receiving the feedback compared to those participating in the control condition.

Decisions about the use of the recommended cutscore can be made based on the 95% confidence interval of the Round 3 cutscores, which was 30.11 to 35.93 for the treatment group and 30.60 to 37.97 for the control group. These bands can serve as estimates of the confidence intervals in cases in which future versions of the test are equated with the
current set (as discussed in Clauser, Harik, et al., 2009; Clauser et al., 2002; Raymond & Reid, 2001), which would mean that the cutscore would be at the same level of difficulty for the equated test versions.

When examining $\hat{\sigma}_{X|I}$ on the probability scale, my study’s estimates ranged from .0276 (in Round 3 of the treatment group) to .0350 (in Round 3 of the control group). Compared to results reported in other studies that treat items as fixed (Brennan & Lockwood, 1980; Clauser, Harik, et al., 2009; Fowell et al., 2008; Verhoeven et al., 2002), these errors are high. With these four studies, the $\hat{\sigma}_{X|I}$ (calculable based on the variance components reported in these studies) were all lower than .0250. Each of these studies, however, investigated variability in cutscores on health or medical licensure tests; typically, judges selected for setting cutscores in these fields are recruited from a homogenous population of judges with specific content or professional expertise. In contrast, judges recruited for education tests are typically recruited from a heterogeneous population that represents several stakeholder groups (Cizek & Bunch, 2007; Hambleton & Pitoniak, 2006; Loomis, 2012). When cutscores are used for making decisions in education contexts, diversity among judges strengthens their validity (Hambleton et al., 2012). Participants in the current study came from a variety of backgrounds rather than a narrow field of expertise, adding to the validity of the standard-setting procedure but resulting in greater likelihood that judges differed in their values and in their beliefs about how much knowledge, skills, and ability is enough to be considered a barely proficient examinee.

The standard errors of the cutscore in this study contribute to the research-base on precision in Angoff methods. Unlike several studies (e.g., Clauser, Harik, et al., 2009;
Clauser et al., 2002; Fehrmann et al., 1991; Fowell et al., 2008; Hurtz & Hertz, 1999; Verhoeven et al., 1999; Verhoeven et al., 2002) reporting or describing their use of the standard error of the cutscore, my study did not allow judges to discuss the items. Scholars (e.g., Geisinger & McCormick, 2010; Hambleton et al., 2012; MacCann & Stanley, 2004; Ricker, 2006) caution against calculating the standard error of the cutscore in rounds after feedback activities, such as discussion, that would cause the judgment data to violate the assumption of independence. Thus, the current study may provide more realistic estimates of the standard error when judges are not subjected to social desirability effects (such as those discussed in Fitzpatrick, 1989).

The data suggest that intrajudge-consistency feedback reduces the standard-errors of the cutscore, \( \hat{d}_{X|I} \), whereas lack of feedback seems to increase these standard errors. In conjunction with the treatment group’s reduced item-judge variance and the increased dependability-index estimates, there is reasonable evidence to assume that intrajudge-consistency feedback improves interjudge consistency. Thus, the feedback appears to strengthen the credibility of the cutscores.

**The Number of Judges Needed**

In addressing Research Question 3, I adopted the same criterion that Hurtz and Hertz (1999) used in their D study to identify how many judges should be included in Angoff procedures. In their study, they looked at eight exams’ standard-setting sessions, which had an average of 7.75 judges (\(SD = 1.67\)) and 170 items (\(SD = 42.97\)) and found that to achieve a standard error of the cutscore, \( \hat{d}_X \), of less than 2.5 raw-score points, 15 judges would be needed on each of the exams. They recommended that Angoff studies in
general have between 10 and 15 judges on the basis that six of the eight exams met this
criterion when the number of judges was set to 10.

Similar to Hurtz and Hertz’s (1999) recommendation, my results suggest that a
minimum of about 10 judges would be needed to achieve the 2.5 raw-score standard-error
criterion. During the final round, the panel receiving intrajudge-consistency feedback
achieved an acceptable degree of precision with nine judges; the panel not receiving this
feedback met the criterion with 11 judges. The D-study plot also suggested that the return
on including more judges after about 10 or 11 appears to diminish. It is likely that other
standard-setting studies with tests measuring similar constructs will find that having
about 10 judges will suffice.

The Round 1-to-3 changes in precision based on Hurtz and Hertz’s (1999) criterion,
suggest that the feedback might prevent a slight loss in precision that is equivalent to the
cost of hiring one judge. Whereas having 10 control-group judges would have sufficed
during Round 1, this number worsened to 11 judges by Round 3. The change for the
treatment group was negligible, with 9 judges needed regardless of the round. This
finding should be interpreted with caution, though, because these scenarios are based on
estimated variance components and the assumption that future judges will be a random
sample from the same population.

The changes in precision can also be examined using the dependability index.
Whereas no predetermined desirable criterion exists for this measure, the traditional .80,
which Raymond and Reid (2001) used as an example for the sake of argument, can be
invoked. To meet a .80 dependability-index criterion, seven control-group judges would
be needed during Round 1; this number worsens to eight control-group judges during
Round 3. The change for the treatment group went from needing five during Round 1 to needing only three judges during Round 3. Thus, for practitioners deciding to use the .80 dependability-index criterion, the intrajudge-consistency feedback appears to be worth about the same cost as hiring at least two more judges. Ultimately, though, decisions about the number of judges to include on a panel should be based not only on data such as those I provide here, but also on policy concerns, such as the stakes of a cutscore and the availability of resources for the standard-setting procedures.

Skeptics might argue that Hurtz and Hertz’s (1999) criterion is dependent on the number of items on the exam. Their 2.50-raw-scale criterion can be translated onto the probability scale: This turns out to be about .0147, which is lower than that observed in the current study with 18 treatment-group judges (.0276). To achieve this criterion with the intrajudge-consistency feedback panel, 64 treatment-group judges would be needed, which is clearly unfeasible. One thing that must be considered, however, is that the judges in Hurtz and Hertz’s procedures had participated in discussion; those in the current study did not. Another point is that the exams in Hurtz and Hertz’s study were for licensure, which do not require broad representation of stakeholder groups; the same cannot be said of standard-setting procedures for education tests (Cizek & Bunch, 2007; Hambleton & Pitoniak, 2006; Hambleton et al., 2012).

Limitations and Suggestions for Future Research

As is common in investigations examining the effect of an intervention for the purpose of making causal inferences, the current study is susceptible, to some extent, to threats to validity. Questions can arise about the degree to which the inferences drawn are valid. And, as with any study, there are limitations.
One question, in this study, is whether participants receiving intrajudge-consistency feedback indeed use the information to improve their judgments about how the BPE would perform on the item or whether they make the adjustments for other reasons. The findings in Clauser, Mee, et al. (2009) suggested that judges tend to mechanically accept the feedback data when they are presented as numerical values. The current study addressed this threat by providing non-numerical data and by explicitly instructing judges to use the feedback to improve their judgment if they believed it helped in this endeavor. Furthermore, the think-aloud protocol data collected during the pilot testing did not reveal that participants were mechanically adopting feedback or that they were systematically changing their item ratings for purposes other than improving their judgments. Still, because a comprehensive think-aloud investigation of the 36 judges was beyond the scope of this study, I cannot state with certainty that the judges used the feedback in the manner that was intended. More extensive verbal protocol research (similar to that conducted by Skorupski & Hambleton, 2005) into what judges are thinking while reviewing and using intrajudge-consistency feedback would provide deeper insight into this issue.

Related to this question is whether there are alternative explanations, beyond those I considered, for judges’ improvements in consistency. I endeavored to address plausible alternative explanations by conducting randomized-group assignment procedures and by examining whether three participant-background characteristics—content-expertise, assessment expertise, and experience in middle- or high-school contexts—explained variability in the model. The randomization resulted in reasonably similar experimental groups, with regard to participant-background features, and I found no evidence
suggesting that the three characteristics predicted changes in consistency indexes. However, even though these three background characteristics have been discussed in the literature as possible explanations for variability in judges’ accuracy, it is conceivable that other, unexplored, judge characteristics contributed to the effects found in this study. Additionally, because I used participants’ self reports to operationalize the three covariates, their scores were susceptible to some degree of construct-irrelevant variance (such as participants’ degree of modesty, beliefs about self efficacy, and so forth). Agencies facilitating standard-setting procedures for high-stakes purposes typically have greater resources for gathering information about participants’ backgrounds and degrees of expertise; with the NAEP, for instance, external indicators of expertise and experience (such as teaching awards, nominations by college officials, and evidence of professional achievements) are used (Loomis, 2012). With these resources, future studies can more thoroughly investigate possible participant background characteristics and any effects they may have on the strength of feedback.

One possible limitation of this study is that the Angoff procedure included only one type of feedback. Including multiple types of feedback would have strengthened the procedural validity of the standard setting procedures. Having only one type of feedback was necessary, however, to ensure the effect of intrajudge consistency was isolated from other types of feedback and to ensure this small-scale study was feasible. One way around this limitation would be to stagger the types of feedback activities across several rounds, as studies investigating conditional $p$-value feedback (Clauser, Harik, et al., 2009; Clauser, Mee, et al., 2009) have done. Such an approach would likely require more than three rounds of rating (as was done in Clauser, Mee, et al., 2009), increasing costs.
Another limitation of this study is that it was not research on an existing, operational, standard-setting procedure; rather, it was on a procedure soliciting cutscores on a low-stakes, somewhat contrived, performance standard. At the same time, this was a strength because it permitted the effect of intrajudge-consistency feedback to be compared to a similar between-round activity. Thus, at the expense of not having strong generalizability to high-stakes standard-setting scenarios, this study has provided evidence about the causality of one type of feedback. Furthermore, because there appears to be no previous experimental research investigating this feedback, this study contributes to the literature and provides evidence-based suggestions to practitioners.

Finally, critics might argue that because the feedback did not result in observable changes in the resulting recommended cutscores, the feedback is of little value. However, as Hambleton, Pitoniak, and Copella (2012) discuss, this is a common occurrence with feedback; according to them, feedback typically serves a psychological purpose; specifically, it enhances the credibility of the recommended cutscore. Furthermore, whereas the feedback in this study did not result in changes in the cutscore, when it is combined with other types of feedback, the result may differ. Future experiments should investigate the causal effects of combined types of standard-setting feedback.

**Conclusion**

The primary purpose of feedback in the Angoff method is to ensure that panelists understand the process and provide their best judgment of the BPE’s performance on each item. When panelists make capricious judgments about examinee performance on test items, the validity of the recommended cutscore is suspect. The intrajudge-consistency index is a measure of the degree to which each panelist is consistent in
making judgments across items. This index will likely be low with panelists making capricious judgments and high with panelists making careful judgments. Item-level intrajudge-consistency feedback alerts judges when they need to more carefully examine their judgment about how the BPE will perform on the item. Until now, evidence that this type of feedback functions in its intended manner has been limited to self-report data in standard-setting procedures that conflate its effects with other types of feedback. The results here provide evidence that this feedback helps panelists make good judgments.

Consumers of standard-setting cutscores use panels’ recommendations to make decisions about examinees. When there are large discrepancies among panelists in their judgments of how much knowledge, skill, or ability is required for a BPE to successfully respond to a set of test questions, there is a broad band of uncertainty in the panel’s recommended cutscore. Interjudge consistency is low. This translates into error in the consumers’ classification decisions and, therefore, places the credibility of the cutscore into question. The results here provide evidence that intrajudge-consistency feedback contributes to improved interjudge consistency, thereby strengthening the credibility of the recommended cutscore.

Conducting standard-setting procedures requires extensive resources. Practitioners must be selective in determining which types of feedback to include, as time and other logistical resources allocated for the procedures are usually limited. For practitioners with access to item-level IRT data, intrajudge-consistency feedback is a worthwhile activity. Logistically, it is convenient in that, unlike other types of feedback such as discussion or rater-location data, intrajudge-consistency feedback can be implemented in sessions attended by a single judge. Additionally, as this study has demonstrated, intrajudge-
consistency feedback can be administered in Excel, a medium that Bunch (2012) has suggested is convenient for judges to enter standard-setting ratings.

Likewise, practitioners must consider how many panelists to recruit, as the cost of hiring and training judges, and of securing the physical resources needed to conduct the procedures, depends on how many participants are involved. The results in this study suggest that practitioners planning to conduct Angoff procedures on multiple-choice tests of vocabulary or other relatively unidimensional construct in the field of education would be wise to include around nine or ten judges when these judges receive intrajudge-consistency feedback. Including more than 11 may result in diminishing returns. With procedures that include more types of feedback and discussion, practitioners can likely achieve acceptable confidence in the cutscore with fewer judges.

This study investigated the effects of item-level intrajudge-consistency feedback, which has long been suggested among standard-setting scholars (e.g., Hambleton et al., 2012; MacCann, 2009; Plake et al., 1991; Reckase & Chen, 2012; van der Linden, 1982) but not experimentally researched and reported in the literature. This study contributes to the literature on feedback in standard setting and provides substantive support for including intrajudge-consistency feedback in Angoff procedures. Researchers and standard-setting practitioners seeking a feedback activity that will likely improve the validity of their procedures now have evidence that may help them in their selection.
APPENDIX A

INSTRUMENT AND PROCEDURE DEVELOPMENT OUTLINES

Recruitment materials and procedures
1. Developed a standardized recruitment email message, drawing from a modeled recruitment letter from the university’s institutional review board and suggestions in the standard-setting literature (Cizek & Bunch, 2007; Raymond & Reid, 2001)
   a. Edited for conciseness, using suggested edits from a senior researcher
2. Developed a standard follow-up correspondence message for
   a. providing participants with the consent form,
   b. arranging the standard-setting session meeting place and time, following suggestions in Cizek and Bunch (2007), and
   c. providing a link to the online background questionnaire
3. Wrote a post-session email message thanking participants and asking if they knew any other potential participants with the desirable qualifications
4. Pilot tested the recruitment messages during the pilot-testing sessions
5. E-mailed messages at appropriate times during main study, adding personalized lines appropriate for specific recipients and lists

Background questionnaire and procedures
1. Identified background information relevant to the standard-setting tasks, as recommended in the standard-setting literature (e.g., Brandon, 2004; Chang et al., 1996; Cizek & Bunch, 2007; Raymond & Reid, 2001; Reckase & Chen, 2012; Verheggen et al., 2008); determined need to investigate participants’ experience and knowledge in the following:
   a. middle- and high-school teaching
   b. college (and university) teaching
   c. language arts teaching
   d. assessment and standard setting
   e. theories of learning reading and vocabulary
2. Drafted items and questionnaire instructions, using questionnaire-writing guidelines in Dillman (2000) and Brown (2001)
3. Edited instructions and item wording, using
   a. suggested edits from a senior researcher knowledgeable in standard setting
   b. think-aloud protocols from the first two pilot-testing participants
4. Prepared the questionnaire for online administration via SurveyMonkey
5. Prepared a method for collecting responses using confidential participant codes rather than participant names
6. Administered the questionnaire to remaining pilot-testing participants
   a. Examined responses
      i. to ensure correct online data collection functioning
      ii. to examine whether participants responded as expected to the questionnaire prompts
iii. to measure how much time participants took in completing the questionnaire
b. Asked participants for their feedback on the questionnaire wording and formatting; prompted for this feedback
   i. on a final question on the questionnaire, and
   ii. in-person at start of standard-setting sessions
c. Made minor changes to the wording and formatting based on pilot-testing feedback
7. Administered to main-study participants prior to their arriving at the standard-setting sessions
8. Quickly reviewed responses immediately prior to the standard-setting session to ensure participants completed the questionnaire
9. Downloaded the responses and calculated descriptive statistics for each item in SAS

Random assignment materials and administration procedures
1. Planned random assignment procedures, consulting Shadish, Cook, and Campbell (2002)
2. Identified strata for random assignment, rand-secondary and rand-other
3. Created an 18-by-3 table comprising randomly generated numbers for each of the two experimental groups within each of the two strata
   a. The first two columns were for each of the two strata
      i. Created a random number in each cell using Excel’s RAND function, ensuring no duplicated numbers were generated
   b. The third column was for experimental group assignment
      i. Entered H (for control group) in the first 9 rows and T (for treatment group) in the last 9 rows
         1) These codes, rather than “C” and “T”, were to avoid participants inadvertently viewing the codes during the session and becoming aware of their experimental group status—the plan was that if participants asked what these meant, they were to be told “H” was for heads and “T” was for tails
   c. Printed the table
4. For each of the two sets of random numbers, printed each number onto its own card
5. Created two bags for containing the cards, one labeled “Rand-secondary” and the other bag labeled “Rand-other”
   a. Placed the cards in their respective bags, verifying all cards were in the correct bag and then shuffling the cards within each bag
6. After each participant completed the background questionnaire, determined which stratum he or she belonged to, depending on the response to the questionnaire item asking about middle- and high-school-teaching experience
   a. Recorded the stratum on the prepared observation log for the standard-setting session
7. At the start of the standard-setting session, had each participant draw a card from assigned stratum bag
a. Looked up the random number on the table to identify which experimental group the participant was assigned
b. Recorded on the observation log each participant’s drawn random number and the resulting assigned experimental group
c. Discarded the card

8. When one group’s bag had only one card left, placed that remaining card in the other bag and had all subsequent participants draw from that bag regardless of their stratum assignment

Orientation sheets
1. Identified the information participants should review at the start of the standard-setting sessions, using suggestions in the standard-setting literature (Cizek & Bunch, 2007; Raymond & Reid, 2001)
2. Determined there was a need to provide participants with the following:
   a. Purpose of the session and role of participants, including how the item ratings would be used
   b. Description of the session they would participate in
   c. Definition of the hypothetical student
   d. Description of the test and a sample test item
   e. Participants’ tasks
      i. Practice with a test item
3. Determined need to standardize orientation in order to make all standard-setting sessions as similar as possible
   a. Decided to use paper-based orientation sheet and pre-scripted verbal instructions when administering the orientation sheets
4. Determined need to make the treatment and control orientations equivalent except in descriptions of between-round tasks
5. Drafted two orientation sheets, one for the treatment and one for the control group
   a. Used exact wording for all text except that required for describing between-round tasks
6. Revised the orientation sheets using
   a. suggested edits from senior researcher knowledgeable in standard setting
   b. notes taken during think-alouds while first two pilot-testing participants read the orientation sheets
7. Added verbal verification questions, included in facilitator protocol checklist, to get verification that participants understood
   a. their tasks,
   b. the definition of the hypothetical student,
   c. what the cutscore means
8. Pilot tested with remaining pilot-testing participants, editing for clarity and to ensure participants correctly defined the hypothetical student and their tasks
9. Prepared final draft of sheets
10. Administered as hard-copy prints during main study at start of standard-setting sessions
**Facilitator protocol checklist**

1. Identified a need to adhere to the same procedures across standard-setting sessions
   a. In order to decrease chance that sessions differed to the extent that such differences would confound any effect that may have been due to the treatment
2. Drafted a list of the procedures I needed to carry out in the study
   a. Consulted Hambleton (2001) for suggestions in planning the procedures
3. Drew from the list of procedures in developing a draft of the checklist
4. Drafted scripted prompts to ensure I asked participants the same questions across standard-setting sessions
5. Reviewed the checklist with a senior researcher familiar with standard-setting research, then revised
6. Used the checklist during the pilot-testing sessions, making revisions immediately after the sessions by reviewing observation log notes that recorded participants’ questions and think-aloud data.
   a. Revisions included the wording of the scripted prompts and the ordering of the steps in the procedure
   b. Fine-tuned the checklist during late pilot-testing sessions
7. Used the checklist for the operational sessions
   a. Reviewed the checklist prior to each session
   b. Followed the checklist during the sessions

**Session observation log**

1. Identified a need to document events that occur during the standard-setting sessions
   a. To record participant codes, randomized-assignment results, time of events, questions posed by participants, and descriptions of any unusual events that occur during the sessions
   b. To document any potential threats to validity
2. Used the logs during pilot testing for recording think-aloud notes and other data that could pose threats to validity
3. Edited the logs to facilitate recording of data
   a. Added fields to match the edited procedures in the facilitator protocol checklist
   b. Resized some fields to accommodate note-taking
4. Prepared an observation log for each participant prior to the standard-setting sessions

**Standard-setting entry forms**

1. Identified a need to administer the standard-setting forms in a way that would allow item-level intrajudge consistency to be immediately calculated in order to provide feedback to treatment group participants
   a. Determined that Excel would be appropriate, a medium that is also discussed in Bunch (2012)
2. Drafted practice-session form
a. Imported 10 practice set items from Vocabulary Size Test items into Excel
b. Developed Excel calculations to provide instant intrajudge consistency feedback, based on van der Linden’s (1982) calculations, to the treatment group
c. Developed equivalent calculations for providing between-round instructions to the control group
d. Developed Excel macros and designed the page appearance to minimize distractions and mistakes by preventing participants from viewing or editing menus, tabs, and hidden sheets
e. Revised for copyedits and formatting
f. Pilot tested with simulated data and verified that the calculations were accurate using a separate Excel sheet

3. Developed 76-item operational session sheet using the practice-set sheet format
a. Pilot tested with simulated data and verified accuracy with calculations on a separate Excel sheet

4. Conducted early pilot testing of the practice- and the 76-item operational-session forms with one pilot-testing participant
a. Think-aloud protocol revealed fatigue effects and that the task exceeded the allotted three hours for the session
b. Determined need to shorten the operational-session sheet
c. Identified and corrected copyedit and formatting errors

5. Revised the practice- and operational-session forms
a. Retained 50 of the 76 items, to shorten the operational-session form

6. Administered the forms to remaining pilot-testing participants, collecting think-aloud protocol data to identify minor edits
7. Revised based on think-aloud data
8. Prepared final operational version
9. Administered to participants

Post-session questionnaires
1. Identified a need, during early pilot testing, for a standardized method to verify that participants were paying adequate attention to essential information in the practice standard-setting form
2. Drafted a post-practice session questionnaire midway through the pilot testing
a. Followed questionnaire-item guidelines in Dillman (2000)
3. Also drafted a post-operational session questionnaire asking the same questions
4. Pilot tested the forms along with the standard-setting entry forms
a. Immediately after completing the practice session, participants were asked to complete the brief post-session questionnaire
b. Participants were asked to verbalize their thoughts as they completed the form, following the think-aloud protocol
c. The same procedure was used with the post-operational session
5. Discovered that a value of the form after the practice session was in identifying participants who might require more instruction and practice before they rated items in the operational task
6. Further revised the questionnaires, using think-aloud protocol data to guide editing decisions
7. Verified their functionality in pilot-testing sessions
8. Prepared final draft for use in main study
APPENDIX B

FACILITATOR PROTOCOL CHECKLIST

Before the session

1. After hearing from the participant and setting up a time and place to meet, email them the consent form. Instruct them to review the form and to complete the background questionnaire. Use the following as a guide in writing the email:

   Dear <participant>,
   Thanks again for your interest in participating in my study. I reserved <room location> for our <time> meeting <day of week and date>. Attached is a consent form for you to review before the session (I will also have a hard-copy of this for you to sign). After you review the form, decide whether you give your consent. If you agree to participate, please complete the online background questionnaire. This asks for background information that might be relevant to the study. It should take about 5 minutes to complete and would need to be done in one sitting.

   On the first page of the questionnaire, it will ask you to enter a code (this code is unique to you). Please enter this code: <code>.

   Here's the link to the questionnaire: https<link>

   (If clicking on the link does not work, you might have to copy and paste it into your browser.)

   Looking forward to it!
   <my name and contact information>

2. One or two days before the meeting, send the participant a reminder email, reconfirming our appointment time and location. Include a map to the location, information about parking, and whether the room will have air conditioning.

3. Prepare the standard-setting Excel files
   a. Prepare both the control and treatment data collection files by labeling them with the participant’s ID code, followed by the letter $H$ for the control files and $T$ for the treatment files.
   b. Place these in a folder labeled with the participant’s ID and copy this folder to a USB memory stick and to each of the PCs that will be used in the session.

4. Prepare the PCs.
   a. Count how many participants will be at the session.
   b. Make sure each machine has the following: a keyboard, mouse, monitor, monitor cable, power cable.
   c. At least 30 minutes before the session, haul the PCs and an extension power cord to the session room.
   d. Assemble and power up each machine.
   e. On each machine, open one of the standard-setting Excel files to make sure it displays correctly. Close it.

5. Prepare the observation log for each participant and the random-number bags.
   a. Print an observation log for each participant and enter the participant’s ID code.
   b. Print the participant’s name on a Post-it note and temporarily tack it to the observation log.
c. Check the participant’s response to Question 2 in the online questionnaire. If the response is *yes*, assign the participant to the stratum labeled *Secondary*, if the response is *no*, assign them to the *Other* group.

d. Carry the experimental group-assignment bags to the session room.

6. Prepare a hard-copy folder containing the following:
   a. this protocol,
   b. hard-copies of the consent form,
   c. the observation logs,
   d. a hard-copy of the blank background questionnaire,
   e. the orientation sheets for both the control and treatment groups,
   f. the post-practice questionnaire for both the control and treatment groups,
   g. the post-actual session questionnaire for both the control and treatment groups,
   h. the envelope with the $20 for their participation.

7. Prepare the refreshments by carrying the snacks, drinks, napkins, and paper plates to the room.

8. For each participant, bring a pen and a pair of earmuffs to the room.

9. Place a consent form at each computer station.

**During the session**

10. Greet each participant as they arrive and direct them to select a computer to sit at.
    a. Ask them to read and complete the consent form if they have not done so already.
    b. Collect the consent forms.
    c. Direct their attention to the refreshments and the location of the nearest restroom.
    d. Tell them that they may take a break at any time if they wish to do so and that if they feel uncomfortable at any time, they may discontinue participating.
    e. Explain that this session will take up to three hours, that the first hour will be orientation and practice, followed by a break. The remaining time will be the actual standard setting procedure. Tell them that they may ask questions at any time, but please do not discuss the activity with other participants.
    f. Tell participants to don the earmuffs if outside noises happen to become distracting.

11. If the participant has not yet completed the online questionnaire, ask them to do so now. Provide the hard-copy if internet connection is unavailable.

12. On each participant observation log, do the following:
    a. Record the date, location, and computer in the observation log.
       i. Indicate if the computer is the small laptop with the external monitor, the XP machine, or if it is the large Vista laptop.
    b. Make a note of any unusual circumstances.

13. At the start of the session, ask the participant if they are comfortable and ready to begin or if they need to take a short break before getting started.

14. Assign the participant to the treatment or control group using the following procedure:
    a. Ask the participant to draw one card from the bag corresponding to the stratum they are in, the *Secondary* or *Other* group.
15. For each participant’s PC, open the practice session file for the experimental group which the participant is in. Make sure the page displayed is the ID sheet, which asks the participant to enter their participant ID code.

16. Orientation sheet
   a. Give the orientation sheet to the participants and ask them to read it very carefully and to ask any questions as they go through it.
   b. After they finish, ask them “Just to make sure you understand your role and tasks, I’m going to ask you to reiterate these things to me.” Then, say:
      i. “So, can you tell me what your role is and what tasks you’ll participate in?”
         1. Address any misconceptions or questions
         2. Tell them that their role is to serve as a standard-setting judge and that their purpose is to give their best judgment about how many of the 100 barely proficient students, that they have in mind, will select the correct answer choice. Tell them that this information will be used to decide where to place the cutscore on the test.
      ii. “Also, just to make sure, can you reiterate back to me the description of the hypothetical students?”
         1. Address any misconceptions or questions.
         2. Tell them that the purpose of conceptualizing 100 students is…
            a. to consider that there will be many types of students with many different characteristics, and that these characteristics are not what we are focusing on. Rather, the one thing they all have in common is this degree of proficiency in their vocabulary knowledge.
      iii. “Just to make sure, can you tell me the difference between Example 1 and Example 2?”
         1. Address any misconceptions or questions.
         2. Tell them that the purpose for looking at these examples is…
            a. to make sure they see that the difficulty of an item might be due to its distractors being similar in meaning to the correct choice, and
            b. to emphasize that it is important to read the entire item before making a judgment.
   c. Point to the agenda on the observation sheet and say again what the schedule will be and how long the practice session take.
      i. Tell them that the practice session will take between 30 and 45 minutes and the actual session will take between 45 and 90 minutes.
      ii. Tell them that we will take a break after the practice session.
      iii. Remind them that they can request a break at any time.
iv. If there are two or more participants, tell them that you will begin the
sessions at the same time and that if one or more of you finish before the
other participant, you can take a break.
v. Ask participants to avoid talking with each other about the items or the
procedures during the session.
d. Ask if anyone has any questions.

17. Practice session
a. Direct the participant to the practice session file on the computer. Tell them that
this is the practice session, that it will take up to 45 minutes. Ask them to
carefully read the instructions and complete the tasks during this practice session
because during the actual session, the instructions will be shorter. Tell them to
please ask questions if any part is unclear.
b. Give them their participant ID code and ask them to enter it on the first page.
c. Allow the participant to continue throughout the practice session.
d. Record the start time of the practice session.
e. After the participant completes the practice session, record the time and ask the
participant to complete the post-practice questionnaire.
   i. If one or more participants finish this section before the others, ask them to
take a break after completing this post-practice questionnaire.
f. After everyone has completed the post-practice questionnaire, ask if they have
any questions or comments.
   i. Record notes in the observation log, if any questions or issues emerge.
g. Ask if they noticed the previous round’s ratings and the answer keys, and if they
knew that they could go back and change their rating, if they believed that this
information improved their judgment.
h. Remind them that the task here is to come to their best judgment about how well
the barely proficient students that they have in mind will perform on each item.
i. Save the practice file, then rename the file, adding “completed” and the date at the
end of the file name. Then, close the file.

18. Break
a. Tell them that the remainder of the session will take up to 90 minutes.
Recommend that the participant take at least a five-minute break.

19. Operational session
a. Open the operational session file on each participant’s computer. Make sure the
page displayed is the ID sheet, which asks the participant to enter their participant
ID code.
b. Ask them to enter the participant ID code on the first screen.
c. Before moving to the next screen, tell them that in this session,
   i. they will rate 50 items in three rounds, instead of just two,
   ii. they will review information about the items twice this time, once after
      Round 1 and again after Round 2,
d. Record the start time of the operational session.
e. Record any questions or issues that might emerge.
f. After the participant completes the operational session, record the time.
g. Save the operational file, then rename the file, adding “completed” and the date at the end of the file name. Then, close the file.

h. Ask the participant to complete the post-actual-session questionnaire.

i. If one or more participants finish before others, ask them to take a break before the final discussion at the end.

j. Ask, “Do you have any comments you would like to share?” and “Do you have any questions?”

i. Record notes in the observation log, if any questions or issues emerge.

☐ 20. Thank the participants for their time and hand each the envelope containing the money for their participation.

☐ 21. Return everything to my office.
   a. Copy the completed files to the USB memory stick.
   b. Turn off the computers and break down.
   c. Place all completed hard-copy forms in the folder labeled completed.
   d. Upload the completed files to my office machine and to the server for backup.
APPENDIX C
SESSON OBSERVATION LOG

<table>
<thead>
<tr>
<th>Instrument</th>
<th>Start time</th>
<th>End time</th>
<th>Notes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Background questionnaire</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Orientation</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Practice Session</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Actual session</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Participant ID: | Location:  
Date: | Computer:  
Secondary experience: | RandNum: | Group:  

154
APPENDIX D

ORIENTATION SHEETS

In participants’ copies, for both experimental groups, the title was simply “Orientation”. The experimental group is included in the titles here for clarity.

Treatment Group Orientation

Thank you for agreeing to participate in this research project. In this project, I’ll first ask you to think about a hypothetical student. Then, I’ll ask you to carefully look at a set of test items (multiple-choice questions) and make judgments about these items based on your conceptualization of the hypothetical student.

Your role will be as a judge in a procedure that is similar to what policy makers use to determine where to place cut scores on assessments. These cut scores are also known as performance standards. Based on your judgments about each item, a cut score on the test will be calculated. Your cut score will be combined with that of other judges and eventually used to identify a final cut score.

The purpose of this cut score will be to identify high-school students (in Hawaii) who have barely enough vocabulary to be considered prepared for college reading. High-school seniors (or students in other grades who attempt the test) who score above this cut score will be classified as prepared for the vocabulary demands they will face in college reading. Students who score below this cut score will be classified as not yet prepared.

Here is the agenda for this procedure and other activities. These activities are explained below.

Agenda

1. Background questionnaire (already completed online)
2. Orientation (this document)
   a. Description of each of the following: your role, the purpose of the procedure, the test you’ll be analyzing, and the activities you’ll participate in
3. Practice standard-setting procedure
   a. Orientation in how to navigate the computer screen and enter ratings
   b. Practice Round 1 item ratings
   c. Practice feedback
   d. Practice Round 2 item ratings
4. Standard-setting procedure
   a. Round 1 item ratings
   b. Feedback
   c. Round 2 item ratings
   d. Feedback
   e. Round 3 item ratings
The Vocabulary Test

The test I will ask you to examine is the Vocabulary Size Test. This test was carefully constructed using information from theories of vocabulary learning and reading. Its reliability and validity have been examined and judged to be very high. Each multiple-choice item has a vocabulary word and an example sentence with that word. Here is an example:

Example item
latter: I agree with the latter.
   a. man from the church
   b. last one
   c. reason given
   d. answer

For each test item, there is only one best answer. Here, it is b, last one.

The Hypothetical Student

Consider students living in Hawaii. Now, picture in your mind a high-school senior who has barely enough vocabulary knowledge to be able to read and understand typical college textbooks without having to use a dictionary or other aid. Take a moment to carefully imagine this student’s vocabulary abilities. During today’s activities, we’ll call this the barely proficient student.

After you have a clear image of this student, imagine that there are 100 of these barely proficient students taking the Vocabulary Size Test. Your task will be to estimate how many of these 100 students will get each test item correct.

Look at the example items and estimate how many barely proficient students (from 0 to 100) would be able to correctly answer each one. Write this number in the box to the right of each item. When you examine each item, be sure to read everything in the item, including all four answer choices (the correct answer choice is indicated by the key). Sometimes the answer choices can affect how easy or difficult the item is.

Example 1
see: They saw it.
   a. started
   b. waited for
   c. opened
   d. looked at

   Key: d

Example 2
see: They saw it.
   a. imagined
   b. cut
   c. perceived
   d. looked at

   Key: d
Both examples test knowledge of the word *see* in an example sentence, *They saw it*. Example Item 1 is very easy, and you probably wrote down a very high number—maybe you estimated that 99 or even all 100 barely proficient students would get it right. Example Item 2 has exactly the same word and sentence but different choices, which might make the item more difficult—you probably entered a lower number in the box, such as 90 or 95. The number you write should depend on your conceptualization of the 100 barely proficient students.

I will ask you to rate 50 items on the Vocabulary Size Test. You will rate each item three times, once during Round 1, again in Round 2, and a final rating in Round 3. After Round 1 and Round 2, I will also give you feedback about how consistent you were in your ratings. This consistency is calculated based on the way you rated the full set of 50 items. The purpose of this feedback is to inform you when you may not have carefully examined an item and the way that the barely proficient students will answer it. You can use this feedback to assist you in subsequent rounds of item rating. You will see what the feedback looks like during the practice session.

**Practice**

In a minute, you’ll begin the practice session. This practice session will guide you through the same steps that you’ll later see when you rate the full 50-item test. The practice session will have only 10 items for you to rate, with feedback on 3 items. The actual standard-setting procedure will have 50 items for you to rate, with feedback on 12 items.

**Questions**

If at any time you have any questions or wish to take a break, please let me know. If you feel that you want to discontinue participating in this research, you may do so at any time.
Control Group Orientation

Thank you for agreeing to participate in this research project. In this project, I’ll first ask you to think about a hypothetical student. Then, I’ll ask you to carefully look at a set of test items (multiple-choice questions) and make judgments about these items based on your conceptualization of the hypothetical student.

Your role will be as a judge in a procedure that is similar to what policy makers use to determine where to place cutscores on assessments. These cutscores are also known as performance standards. Based on your judgments about each item, a cutscore on the test will be calculated. Your cutscore will be combined with that of other judges and eventually used to identify a final cutscore.

The purpose of this cutscore will be to identify high-school students (in Hawaii) who have barely enough vocabulary to be considered prepared for college reading. High-school seniors (or students in other grades who attempt the test) who score above this cutscore will be classified as prepared for the vocabulary demands they will face in college reading. Students who score below this cutscore will be classified as not yet prepared.

Here is the agenda for this procedure and other activities. These activities are explained below.

Agenda

1. Background questionnaire (already completed online)
2. Orientation (this document)
   a. Description of each of the following: your role, the purpose of the procedure, the test you’ll be analyzing, and the activities you’ll participate in
3. Practice standard-setting procedure
   a. Orientation in how to navigate the computer screen and enter ratings
   b. Practice Round 1 item ratings
   c. Practice item reexamination
   d. Practice Round 2 item ratings
4. Standard-setting procedure
   a. Round 1 item ratings
   b. Item reexamination
   c. Round 2 item ratings
   d. Item reexamination
   e. Round 3 item ratings
The Vocabulary Test

The test I will ask you to examine is the Vocabulary Size Test. This test was carefully constructed using information from theories of vocabulary learning and reading. Its reliability and validity have been examined and judged to be very high. Each multiple-choice item has a vocabulary word and an example sentence with that word. Here is an example:

Example item
latter: I agree with the latter.
   a. man from the church
   b. last one
   c. reason given
   d. answer

For each test item, there is only one best answer. Here, it is b, last one.

The Hypothetical Student

Consider students living in Hawaii. Now, picture in your mind a high-school senior who has barely enough vocabulary knowledge to be able to read and understand typical college textbooks without having to use a dictionary or other aid. Take a moment to carefully imagine this student’s vocabulary abilities. During today’s activities, we’ll call this the barely proficient student.

After you have a clear image of this student, imagine that there are 100 of these barely proficient students taking the Vocabulary Size Test. Your task will be to estimate how many of these 100 students will get each test item correct.

Look at the example items and estimate how many barely proficient students (from 0 to 100) would be able to correctly answer each one. Write this number in the box to the right of each item. When you examine each item, be sure to read everything in the item, including all four answer choices (the correct answer choice is indicated by the key). Sometimes the answer choices can affect how easy or difficult the item is.

Example 1
see: They saw it.
   a. started
   b. waited for
   c. opened
   d. looked at

Key: d

Example 2
see: They saw it.
   a. imagined
   b. cut
   c. perceived
   d. looked at

Key: d
Both examples test knowledge of the word see in an example sentence, They saw it. Example Item 1 is very easy, and you probably wrote down a very high number—maybe you estimated that 99 or even all 100 barely proficient students would get it right. Example Item 2 has exactly the same word and sentence but different choices, which might make the item more difficult—you probably entered a lower number in the box, such as 90 or 95. The number you write should depend on your conceptualization of the 100 barely proficient students.

I will ask you to rate 50 items on the Vocabulary Size Test. You will rate each item three times, once during Round 1, again in Round 2, and a final rating in Round 3. After Round 1 and Round 2, I will also ask you to reexamine a sample of these items. You will select this sample using your own judgment. Reexamining items might help you to think about how many of the 100 barely proficient students will answer the question correctly. You will see what all this looks like during the practice session.

**Practice**

In a minute, you’ll begin the practice session. This practice session will guide you through the same steps that you’ll later see when you rate the full 50-item test. The practice session will have only 10 items for you to rate, with 3 items for you to select and reexamine. The actual standard-setting procedure will have 50 items for you to rate, with 12 items for you to select and reexamine.

**Questions**

If at any time you have any questions or wish to take a break, please let me know. If you feel that you want to discontinue participating in this research, you may do so at any time.
### Round 1 Item Rating

**Directions:**
Out of 100 high school seniors with barely enough vocabulary to successfully read college textbooks, how many will get the item correct? Enter a number for every item. Remember to look at the item carefully, including all possible answer choices, before you enter your rating. You can change your ratings for an item at any time while you are on this page.

If you have any questions at any time, please ask.

<table>
<thead>
<tr>
<th>Item</th>
<th>Number of barely proficient students</th>
</tr>
</thead>
</table>
| genre: They argued about its **genre**.  
  a. the class it belongs to  
  b. its origin or roots  
  c. the information it contains  
  d. whether it applies to everybody | 61 |
| threshold: They raised the **threshold**.  
  a. flag  
  b. point or line where something changes  
  c. roof inside a building  
  d. cost of borrowing money | 60 |
| vane: The **vane** was broken.  
  a. water pipe  
  b. grey or colored rod inside a pencil  
  c. piece of glass in a window  
  d. pointer that shows the direction of the wind | 50 |
| talon: Just look at those **talons**!  
  a. high points of mountains  
  b. sharp hooks on the feet of a hunting bird  
  c. heavy metal coats to protect against weapons  
  d. people who make fools of themselves without realizing it |
## Feedback on Round 1 Ratings

**Directions:**
You will see up to 12 items below. Your ratings on these items were not consistent with your ratings on the other items in Round 1. Carefully reexamine each item as you review its feedback. Look at the word, the sentence, and the answer choices, and consider how the barely proficient students will answer the item.

Writing comments can help you organize your thinking about the item after you have reviewed the feedback. If you cannot think of anything to write for an item, you can leave that item’s comment box blank.

<table>
<thead>
<tr>
<th>Item number</th>
<th>Key</th>
<th>Item</th>
<th>In Round 1, you said this many would answer it correctly:</th>
<th>Considering the way you rated the full set of items, your rating on this item was:</th>
<th>Considering the way you rated the full set of items, it is likely that:</th>
<th>Your comments:</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>b</td>
<td>threshold: They raised the threshold.</td>
<td>60</td>
<td>very inconsistent</td>
<td>MANY FEWER of the barely proficient students would have selected choice b.</td>
<td>&quot;Threshold of pain&quot; is a pretty common expression that I think most students would have heard or used before. This would eliminate a, c, and d. &quot;Threshold of pain&quot; might remind students of &quot;limit&quot; or &quot;breaking point&quot; which could lead them to b.</td>
</tr>
<tr>
<td>3</td>
<td>d</td>
<td>vane: The vane was broken.</td>
<td>50</td>
<td>very inconsistent</td>
<td>MANY FEWER of the barely proficient students would have selected choice d.</td>
<td>On second thought, I would score this lower. Students could confuse &quot;vane&quot; with &quot;window pane&quot; and choose c. Or they could choose a because water pipes tend to break.</td>
</tr>
<tr>
<td>4</td>
<td>b</td>
<td>talon: Just look at those talons!</td>
<td>50</td>
<td>very inconsistent</td>
<td>MANY MORE of the barely proficient students probably would have selected choice b.</td>
<td></td>
</tr>
</tbody>
</table>

---

162
### Round 2 Item Rating

**Directions:**
Here, you can see the ratings you gave in Round 1 and the feedback on your Round 1 ratings. As before, the judgment you are making is your answer to this question: Out of 100 high school seniors with barely enough vocabulary to successfully read college textbooks, how many will get the item correct? Your ratings can differ from the ratings you gave in Round 1. Give your best judgment and remember to look at all parts of the item, not just the word. You can change your ratings for an item at any time while you are on this page.

<table>
<thead>
<tr>
<th>Item</th>
<th>Number of barely proficient students</th>
<th>Feedback on Round 1 ratings</th>
</tr>
</thead>
</table>
| 1. **genre:** They argued about its **genre**.  
   a. the class it belongs to  
   b. its origin or roots  
   c. the information it contains  
   d. whether it applies to everybody | **60** | Your rating on this item in Round 1 was very inconsistent. It is likely that MANY FEWER of the barely proficient students will get this item correct than what you had specified in Round 1. Your comments: "Threshold of pain" is a pretty common expression that I think most students would have heard or used before. This would eliminate a, c, and d. "Threshold of pain" might remind students of "limit" or "breaking point" which could lead them to b. |
| 2. **threshold:** They raised the **threshold**.  
   a. flag  
   b. point or line where something changes  
   c. roof inside a building  
   d. cost of borrowing money | **55** | Your rating on this item in Round 1 was very inconsistent. It is likely that MANY FEWER of the barely proficient students will get this item correct than what you had specified in Round 1. Your comments: On second thought, I would score this lower. Students could confuse "vane" with "window pane" and choose c. Or they could choose a because water pipes tend to break. |
| 3. **vane:** The **vane** was broken.  
   a. water pipe  
   b. grey or colored rod inside a pencil  
   c. piece of glass in a window  
   d. pointer that shows the direction of the wind | **40** | Your rating on this item in Round 1 was very inconsistent. It is likely that MANY FEWER of the barely proficient students will get this item correct than what you had specified in Round 1. Your comments: On second thought, I would score this lower. Students could confuse "vane" with "window pane" and choose c. Or they could choose a because water pipes tend to break. |
| 4. **talus:** Just look at those **talus**!  
   a. high points of mountains  
   b. sharp hooks on the feet of a hunting bird  
   c. heavy metal coats to protect against weapons  
   d. people who make fools of themselves without realizing it | **60** | Your rating on this item in Round 1 was very inconsistent. It is likely that MANY MORE of the barely proficient students will get this item correct than what you had specified in Round 1. Your comments: (no comment entered) |
## Feedback on Round 2 Ratings

**Directions:**

You will see up to 12 items below. Your ratings on these items were not consistent with your ratings on the other items in Round 2. Carefully reexamine each item as you review its feedback. Look at the word, the sentence, and the answer choices, and consider how the barely proficient students will answer the item.

Writing comments can help you organize your thinking about the item after you have reviewed the feedback. If you cannot think of anything to write for an item, you can leave that item’s comment box blank.

<table>
<thead>
<tr>
<th>Item number</th>
<th>Key</th>
<th>Item</th>
<th>In Round 2, you said this many would answer it correctly:</th>
<th>Considering the way you rated the full set of items, your rating on this item was:</th>
<th>Considering the way you rated the full set of items, it is likely that:</th>
</tr>
</thead>
<tbody>
<tr>
<td>2 b</td>
<td></td>
<td>threshold: They raised the threshold.</td>
<td>a. flag</td>
<td>55 somewhat inconsistent</td>
<td>FEWER of the barely proficient students would have selected choice b.</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>b. point or line where something changes</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>c. roof inside a building</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>d. cost of borrowing money</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Your comments:</td>
<td>Now that I look closer at the context provided by the example sentence I see that a, c, and d are more plausible choices, especially for students who have very little or no familiarity with the word.</td>
<td></td>
<td></td>
</tr>
<tr>
<td>4 b</td>
<td></td>
<td>talon: Just look at those talons!</td>
<td>a. high points of mountains</td>
<td>60 somewhat inconsistent</td>
<td>MORE of the barely proficient students would have selected choice b.</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>b. sharp hooks on the feet of a hunting bird</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>c. heavy metal coats to protect against weapons</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>d. people who make fools of themselves without realizing it</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Your comments:</td>
<td>I probably have an inconsistently low score for this item because the word &quot;talon&quot; is not part of my own active vocabulary. I know the word, but I never really use it.</td>
<td></td>
<td></td>
</tr>
<tr>
<td>7 c</td>
<td></td>
<td>dilapidated: The house was dilapidated.</td>
<td>a. completely knocked down</td>
<td>45 somewhat inconsistent</td>
<td>MORE of the barely proficient students probably would have selected choice c.</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>b. visited by the spirit of a dead person</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>c. in very poor condition</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>d. not lived in</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Your comments:</td>
<td>I think that students may be more likely to recognize/comprehend this word if they heard it in speaking, but in writing it might not register. Also, students could be lead to answer d because dilapidated houses are often abandoned/not lived in.</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
### Round 3 Item Rating

**Directions:**
This is the final round. Here, you can see the ratings you gave in Round 2 and the feedback on your Round 2 ratings. As before, the judgment you are making is your answer to this question: Out of 100 high school seniors with barely enough vocabulary to successfully read college textbooks, how many will get the item correct?

<table>
<thead>
<tr>
<th>Item</th>
<th>Number of barely proficient students</th>
<th>Feedback on Round 2 ratings</th>
</tr>
</thead>
</table>
| 1    | **genre:** They argued about its genre.  
  a. the class it belongs to  
  b. its origin or roots  
  c. the information it contains  
  d. whether it applies to everybody | key: a  
  Round 2: 60 | Your rating on this item in Round 2 was somewhat inconsistent. It is likely that FEWER of the barely proficient students will get this item correct than what you had specified in Round 2.  
Your comments: Now that I look closer at the context provided by the example sentence I see that a, c, and d are more plausible choices, especially for students who have very little or no familiarity with the word.  
  |  |
| 2    | **threshold:** They raised the threshold.  
  a. flag  
  b. point or line where something changes  
  c. roof inside a building  
  d. cost of borrowing money | key: b  
  Round 2: 55 |  |
| 3    | **vane:** The vane was broken.  
  a. water pipe  
  b. grey or colored rod inside a pencil  
  c. piece of glass in a window  
  d. pointer that shows the direction of the wind | key: d  
  Round 2: 40 | Your rating on this item in Round 2 was somewhat inconsistent. It is likely that MORE of the barely proficient students will get this item correct than what you had specified in Round 2.  
Your comments: I probably have an inconsistently low score for this item because the word ‘vane’ is not part of my own active vocabulary. I know the word, but I never really use it.  
  |  |
| 4    | **talon:** Just look at those talons!  
  a. high points of mountains  
  b. sharp hooks on the feet of a hunting bird  
  c. heavy metal coats to protect against weapons  
  d. people who make fools of themselves without realizing it | key: b  
  Round 2: 60 |  |
Your session is complete. Thank you.
Please ask the facilitator for assistance.
### Round 1 Item Rating

**Directions:**
Out of 100 high school seniors with barely enough vocabulary to successfully read college textbooks, how many will get the item correct? Enter a number for every item. Remember to look at the item carefully, including all possible answer choices, before you enter your rating. You can change your ratings for an item at any time while you are on this page.

If you have any questions at any time, please ask.

<table>
<thead>
<tr>
<th>Item</th>
<th>Number of barely proficient students</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td></td>
</tr>
</tbody>
</table>

### Example Entries
Example entries (from one of the participants) are included in these excerpts. The total number of items provided in each round is 50. The number of items in the reexamination sheets is 12.

Round 1–3 rating tasks are identical to those in the treatment group. Only the first three to five items per Excel sheet are shown.
### Item Selection after Round 1

**Directions:**
Select exactly 12 items which you will more closely analyze in the next section. These might be items which you are unsure about or which you have difficulty imagining the way a barely proficient high-school senior will respond. You can see the ratings you gave during the previous round (Round 1).

To select an item for closer examination, type a Y in the box. Leave the other items' boxes blank (or type N).

<table>
<thead>
<tr>
<th>Item</th>
<th>You have selected 12 out of 12 items.</th>
</tr>
</thead>
<tbody>
<tr>
<td>4</td>
<td><strong>talon</strong>: Just look at those <em>talons</em>!</td>
</tr>
<tr>
<td></td>
<td>a. high points of mountains</td>
</tr>
<tr>
<td></td>
<td>b. sharp hooks on the feet of a hunting bird</td>
</tr>
<tr>
<td></td>
<td>c. heavy metal coats to protect against weapons</td>
</tr>
<tr>
<td></td>
<td>d. people who make fools of themselves without realizing it</td>
</tr>
<tr>
<td></td>
<td>key: b</td>
</tr>
<tr>
<td></td>
<td>Round 1: 60</td>
</tr>
<tr>
<td>5</td>
<td><strong>squelch</strong>: They <em>squelched past</em>.</td>
</tr>
<tr>
<td></td>
<td>a. drove past very fast making a loud high sound with the wheels</td>
</tr>
<tr>
<td></td>
<td>b. marched past with their weapons raised</td>
</tr>
<tr>
<td></td>
<td>c. walked past making a sucking noise because of water</td>
</tr>
<tr>
<td></td>
<td>d. danced past weaving in and out in patterns</td>
</tr>
<tr>
<td></td>
<td>key: c</td>
</tr>
<tr>
<td></td>
<td>Round 1: 60</td>
</tr>
<tr>
<td>6</td>
<td><strong>demography</strong>: This book is about <em>demography</em>.</td>
</tr>
<tr>
<td></td>
<td>a. the study of patterns of land use</td>
</tr>
<tr>
<td></td>
<td>b. the study of the use of pictures to show facts about numbers</td>
</tr>
<tr>
<td></td>
<td>c. the study of the movement of water</td>
</tr>
<tr>
<td></td>
<td>d. the study of population</td>
</tr>
<tr>
<td></td>
<td>key: d</td>
</tr>
<tr>
<td></td>
<td>Round 1: 40</td>
</tr>
</tbody>
</table>
## Reexamination of Round 1 Item Ratings

**Directions:**
You selected these 12 items. Reexamine each item and write comments about it. Look at the word, the sentence, and the answer choices, and consider how the barely proficient students will answer the item. If you do not have anything to comment on about an item, you can leave that item's box blank and move to the next item.

Remember to use the Tab key (or to click outside the box) to get out of the box. To go back and add to what you wrote in a previous box, double-click into the box.

<table>
<thead>
<tr>
<th>Item #</th>
<th>Key</th>
<th>Item</th>
<th>In Round 1, you said this many would answer it correctly:</th>
<th>Your comments</th>
</tr>
</thead>
<tbody>
<tr>
<td>6</td>
<td>d</td>
<td>demography: This book is about demography.</td>
<td>40</td>
<td>I think students will choose A or D. I think they might be more familiar with &quot;geography,&quot; and perhaps think they are similar, and choose D. Or, recognize they might be different, and choose A. And then I subtracted 10 from half right (50) and got my answer (for those that are guessing more or not reading carefully and choose B or C). But, I think I'd increase my score a bit as &quot;demo&quot; doesn't sound at all like water or pictures and so I don't think those would be chosen wrong very frequently.</td>
</tr>
<tr>
<td>11</td>
<td>c</td>
<td>emit: The sound was emitted by the machine.</td>
<td>85</td>
<td>I thought, even if you have no idea what emit means, you can figure it out from the answer choices. What machine covers up sound? Students are probably only familiar with photocopies and I don't hear the word &quot;emit&quot; used with them. But - 85 for a correct answer seems high - but, to me, the other choices just seem very unlikely to be picked as good options. I choose this one to reflect on because I think I generally give lower scores, and this one was uncharacteristically high for me on a hard word</td>
</tr>
<tr>
<td>16</td>
<td>e</td>
<td>acquisitive: He is very acquisitive.</td>
<td>20</td>
<td>I choose a very low score as I think it's a hard word - and student might think it's similar to &quot;inquisitive&quot; - a word they are probably much more familiar with. I choose less than 25% (a guess) as I think more students will choose A than random guessing.</td>
</tr>
</tbody>
</table>
### Round 2 Item Rating

**Directions:**
Here, you can see the ratings you gave in Round 1 and the comments you wrote about your Round 1 ratings. As before, the judgment you are making is your answer to this question: Out of 100 high school seniors with barely enough vocabulary to successfully read college textbooks, how many will get the item correct? Your ratings can differ from the ratings you gave in Round 1. Give your best judgment and remember to look at all parts of the item, not just the word. You can change your ratings for an item at any time while you are on this page.

<table>
<thead>
<tr>
<th>Item</th>
<th>Number of barely proficient students</th>
<th>Your comments after Round 1</th>
</tr>
</thead>
<tbody>
<tr>
<td>3</td>
<td>vane: The vane was broken.</td>
<td></td>
</tr>
<tr>
<td></td>
<td>a. water pipe</td>
<td></td>
</tr>
<tr>
<td></td>
<td>b. grey or colored rod inside a pencil</td>
<td></td>
</tr>
<tr>
<td></td>
<td>c. piece of glass in a window</td>
<td>key: d</td>
</tr>
<tr>
<td></td>
<td>d. pointer that shows the direction of the wind</td>
<td>Round 1: 60</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>talons: Just look at those talons!</td>
<td></td>
</tr>
<tr>
<td></td>
<td>a. high points of mountains</td>
<td></td>
</tr>
<tr>
<td></td>
<td>b. sharp looks on the feet of a hunting bird</td>
<td>key: b</td>
</tr>
<tr>
<td></td>
<td>c. heavy metal coats to protect against weapons</td>
<td>Round 1: 60</td>
</tr>
<tr>
<td></td>
<td>d. people who make fools of themselves without realizing it</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>squeak: They squeaked past.</td>
<td></td>
</tr>
<tr>
<td></td>
<td>a. drove past very fast making a loud high sound with the wheels</td>
<td>key: c</td>
</tr>
<tr>
<td></td>
<td>b. marched past with their weapons raised</td>
<td></td>
</tr>
<tr>
<td></td>
<td>c. walked past making a sucking noise because of water</td>
<td></td>
</tr>
<tr>
<td></td>
<td>d. danced past weaving in and out in patterns</td>
<td>Round 1: 60</td>
</tr>
<tr>
<td>6</td>
<td>demography: This book is about demography.</td>
<td></td>
</tr>
<tr>
<td></td>
<td>a. the study of patterns of land use</td>
<td></td>
</tr>
<tr>
<td></td>
<td>b. the study of the use of pictures to show facts about numbers</td>
<td>key: d</td>
</tr>
<tr>
<td></td>
<td>c. the study of the movement of water</td>
<td></td>
</tr>
<tr>
<td></td>
<td>d. the study of population</td>
<td>Round 1: 60</td>
</tr>
</tbody>
</table>

I think students will choose A or D. I think they might be more familiar with "geography", and perhaps think they are similar, and choose D. Or, recognize they might be different, and choose A. And then I subtracted 10 from half right (50) and got my answer (for those that are guessing more or not reading carefully and choose B or C). But, I think I'd increase my score a bit as "demo" doesn't sound at all like water or pictures and so I don't think these would be chosen wrong very frequently.
# Item Selection after Round 2

**Directions:**
Select exactly 12 items which you will more closely analyze in the next section. These items do not have to be the same as those you selected before. You can see the ratings you gave during the previous round (Round 2).

To select an item for closer examination, type a Y in the box. Leave the other items' boxes blank (or type N).

<table>
<thead>
<tr>
<th>Item</th>
<th>You have selected 12 out of 12 items.</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td><strong>genre:</strong> They argued about its <strong>genre</strong>.</td>
</tr>
<tr>
<td></td>
<td>a. the class it belongs to</td>
</tr>
<tr>
<td></td>
<td>b. its origin or roots</td>
</tr>
<tr>
<td></td>
<td>c. the information it contains</td>
</tr>
<tr>
<td></td>
<td>d. whether it applies to everybody</td>
</tr>
<tr>
<td></td>
<td>key: a</td>
</tr>
<tr>
<td></td>
<td>Round 2: 70</td>
</tr>
<tr>
<td></td>
<td>Y</td>
</tr>
<tr>
<td>2</td>
<td><strong>threshold:</strong> They raised the <strong>threshold</strong>.</td>
</tr>
<tr>
<td></td>
<td>a. flag</td>
</tr>
<tr>
<td></td>
<td>b. point or line where something changes</td>
</tr>
<tr>
<td></td>
<td>c. roof inside a building</td>
</tr>
<tr>
<td></td>
<td>d. cost of borrowing money</td>
</tr>
<tr>
<td></td>
<td>key: b</td>
</tr>
<tr>
<td></td>
<td>Round 2: 75</td>
</tr>
<tr>
<td>3</td>
<td><strong>vane:</strong> The <strong>vane</strong> was broken.</td>
</tr>
<tr>
<td></td>
<td>a. water pipe</td>
</tr>
<tr>
<td></td>
<td>b. grey or colored rod inside a pencil</td>
</tr>
<tr>
<td></td>
<td>c. piece of glass in a window</td>
</tr>
<tr>
<td></td>
<td>d. pointer that shows the direction of the wind</td>
</tr>
<tr>
<td></td>
<td>key: d</td>
</tr>
<tr>
<td></td>
<td>Round 2: 60</td>
</tr>
</tbody>
</table>
Reexamination of Round 2 Item Ratings

**Directions:**
You selected these 12 items. Reexamine each item and write comments about it. Look at the word, the sentence, and the answer choices, and consider how the barely proficient students will answer the item. If you do not have anything to comment on about an item, you can leave that item's box blank and move to the next item.

After you have entered your comments, you will rate the 50 items in the final round.

<table>
<thead>
<tr>
<th>Item #</th>
<th>Key</th>
<th>Item</th>
<th>In Round 1, you said this many would answer it correctly:</th>
<th>Your comments</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>a</td>
<td>genre: They argued about its genre.</td>
<td>70</td>
<td>I decided to reflect on this one because on one hand I think it's a common H5 word - more so even than in &quot;real-life&quot; because of the emphasis on test-taking of classifying a written piece on its genre - or other characteristics. But... maybe students might get confused between A and B as both speak to underlying causes - either category or basis. Which is why I didn't rate it as high as I might have initially.</td>
</tr>
<tr>
<td>2</td>
<td>d</td>
<td>vane: The vane was broken.</td>
<td>60</td>
<td>At first I thought it was more obvious. But weather vanes are really really rare nowadays. I don't think I've ever even seen one in Hawaii. A student might choose D because it's the only answer that makes sense - but it they really have no idea what a weather vane is they might think it's another word for either A, B, or C.</td>
</tr>
<tr>
<td>3</td>
<td>b</td>
<td>talon: Just look at those talons!</td>
<td>80</td>
<td>The way the sentence is worded, it's like a &quot;wow - look at that&quot;. And I don't think students would think that would apply to &quot;C&quot;, or even &quot;D&quot;. Leaving &quot;A&quot; and &quot;B&quot;: if they don't know talon but think it's like titan - and applies to size, some might choose A, but I think because talons are associated with &quot;fierce&quot; birds profiled on animal planet and the like, I thought even though not a common word there might be some familiarity.</td>
</tr>
</tbody>
</table>
## Round 3 Item Rating

**Directions:**
This is the final round. Here, you can see the ratings you gave in Round 2 and the comments you wrote about the items you selected from Round 2. As before, the judgment you are making is your answer to this question: Out of 100 high school seniors with barely enough vocabulary to successfully read college textbooks, how many will get the item correct?

<table>
<thead>
<tr>
<th>Item</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. genre: They argued about its genre.</td>
</tr>
<tr>
<td>a. the class it belongs to</td>
</tr>
<tr>
<td>b. its origin or roots</td>
</tr>
<tr>
<td>c. the information it contains</td>
</tr>
<tr>
<td>d. whether it applies to everybody</td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td>70</td>
</tr>
<tr>
<td>2. threshold: They raised the threshold.</td>
</tr>
<tr>
<td>a. flag</td>
</tr>
<tr>
<td>b. point or line where something changes</td>
</tr>
<tr>
<td>c. roof inside a building</td>
</tr>
<tr>
<td>d. cost of borrowing money</td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td>3. vane: The vane was broken.</td>
</tr>
<tr>
<td>a. water pipe</td>
</tr>
<tr>
<td>b. grey or colored rod inside a pencil</td>
</tr>
<tr>
<td>c. piece of glass in a window</td>
</tr>
<tr>
<td>d. pointer that shows the direction of the wind</td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td>4. talon: Just look at those talons!</td>
</tr>
<tr>
<td>a. high points of mountains</td>
</tr>
<tr>
<td>b. sharp hooks on the feet of a hunting bird</td>
</tr>
<tr>
<td>c. heavy metal coats to protect against weapons</td>
</tr>
<tr>
<td>d. people who make fools of themselves without realizing it</td>
</tr>
<tr>
<td></td>
</tr>
</tbody>
</table>

**Note:**
- The table above shows the number of barely proficient students who rated each item and the comments made by the judge after Round 2.
Your session is complete. Thank you.
Please ask the facilitator for assistance.
APPENDIX F

POST-SESSION QUESTIONNAIRE

ID code: (entered by the facilitator)

When making your judgment on items during each of the two practice rounds, to what extent did you pay attention to each of the following components? Circle the number on the 0 to 10 scale, where 0 means you paid no attention to that component during that round, and 10 means you paid complete attention to that component in that round.

<table>
<thead>
<tr>
<th>Component</th>
<th>No attention</th>
<th>Complete attention</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. The instructions about how to rate items during the round</td>
<td></td>
<td></td>
</tr>
<tr>
<td>during Round 1</td>
<td>0 1 2 3 4 5 6 7 8 9 10</td>
<td></td>
</tr>
<tr>
<td>during Round 2</td>
<td>0 1 2 3 4 5 6 7 8 9 10</td>
<td></td>
</tr>
<tr>
<td>2. The vocabulary word (e.g., “tinkle”)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>during Round 1</td>
<td>0 1 2 3 4 5 6 7 8 9 10</td>
<td></td>
</tr>
<tr>
<td>during Round 2</td>
<td>0 1 2 3 4 5 6 7 8 9 10</td>
<td></td>
</tr>
<tr>
<td>3. The sentence the word was in (e.g., “It tinkled.”)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>during Round 1</td>
<td>0 1 2 3 4 5 6 7 8 9 10</td>
<td></td>
</tr>
<tr>
<td>during Round 2</td>
<td>0 1 2 3 4 5 6 7 8 9 10</td>
<td></td>
</tr>
<tr>
<td>4. The incorrect answer choices</td>
<td></td>
<td></td>
</tr>
<tr>
<td>during Round 1</td>
<td>0 1 2 3 4 5 6 7 8 9 10</td>
<td></td>
</tr>
<tr>
<td>during Round 2</td>
<td>0 1 2 3 4 5 6 7 8 9 10</td>
<td></td>
</tr>
<tr>
<td>5. The correct answer choice (indicated by the “key:”)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>during Round 1</td>
<td>0 1 2 3 4 5 6 7 8 9 10</td>
<td></td>
</tr>
<tr>
<td>during Round 2</td>
<td>0 1 2 3 4 5 6 7 8 9 10</td>
<td></td>
</tr>
<tr>
<td>6. The feedback (shown next to the items during Round 2)</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>0 1 2 3 4 5 6 7 8 9 10</td>
<td></td>
</tr>
<tr>
<td>7. Your Round 1 ratings (shown next to the items during Round 2)</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>0 1 2 3 4 5 6 7 8 9 10</td>
<td></td>
</tr>
</tbody>
</table>
APPENDIX G

VERBAL PROTOCOL DIRECTIONS

Thank you for participating in the beginning stages of this project. Today, I’m going to ask you to do a think aloud. The purpose of this think-aloud activity is to see whether the instruments are working as expected and to see what changes, if any, should be made. In this think-aloud, please say aloud EVERYTHING you are thinking while you are reading and while you are completing the tasks in each instrument. Please do NOT stop talking. Talk CONSTANTLY. Also, do not try to plan what to say before you talk. Instead, just talk. Please do not talk to me or with me. Instead, imagine that you are in the room alone. If you are silent for too long, I will ask you to talk. Does this make sense?

To start, I want to warm-up so that you can get used to talking aloud without stopping.

1) First, solve this problem while you think aloud: What is the answer to this problem? What is 12 X 30?

2) How many windows are there in the last two places that you have lived?

3) Read these sentences aloud and think aloud while you are reading to show what you are thinking while you are reading:

   I don’t need a husband or kids. I want a dog.
   Dogs are perfect! They don’t complain.
   They’re dílis to you. And, they care about you and give you love. They even protect you from dangerous people.

Do you have any questions about how to think aloud?

Note. The word dílis is not English (it is loyal in Gaelic), but was included in order to get the participant used to stopping and verbalize that they did not recognize this word.
Table H.1

*Group Assignment Probabilities in the Secondary-experience Stratum*

<table>
<thead>
<tr>
<th>Session</th>
<th>Number of participants</th>
<th>Number of cards available</th>
<th>Probability of selecting group card</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Control</td>
<td>Treatment</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>9</td>
<td>9</td>
</tr>
<tr>
<td>3</td>
<td>1</td>
<td>9</td>
<td>8</td>
</tr>
<tr>
<td>4</td>
<td>1</td>
<td>8</td>
<td>8</td>
</tr>
<tr>
<td>5</td>
<td>1</td>
<td>7</td>
<td>8</td>
</tr>
<tr>
<td>6</td>
<td>2</td>
<td>6</td>
<td>8</td>
</tr>
<tr>
<td>7</td>
<td>1</td>
<td>6</td>
<td>6</td>
</tr>
<tr>
<td>8</td>
<td>2</td>
<td>5</td>
<td>6</td>
</tr>
<tr>
<td>9</td>
<td>1</td>
<td>3</td>
<td>6</td>
</tr>
<tr>
<td>10</td>
<td>1</td>
<td>3</td>
<td>5</td>
</tr>
<tr>
<td>11</td>
<td>3</td>
<td>3</td>
<td>4</td>
</tr>
<tr>
<td>12</td>
<td>2</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>13\textsuperscript{b}</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>

*Note.* There was a total of 23 sessions. In Session 2, none of the participants had secondary-teaching experience; thus, this session is absent from this table. After Session 13, there were no remaining cards in this bag; participants in these later sessions drew from the rand-other bag regardless of whether they had secondary-teaching experience.\textsuperscript{a}

*These cards were in the rand-secondary bag. \textsuperscript{b}Session 13 was the last session in which participants with secondary experience selected from this bag. Because there was only one card remaining after Session 13, which would have required the Session 14 participant to select a treatment card, this remaining card was placed in the rand-other bag.*
Table H.2

*Group Assignment Probabilities in the Other Stratum*

<table>
<thead>
<tr>
<th>Session</th>
<th>Number of participants</th>
<th>Number of cards available&lt;sup&gt;a&lt;/sup&gt;</th>
<th>Probability of selecting group card</th>
<th>Control</th>
<th>Treatment</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>1</td>
<td>9</td>
<td>.50</td>
<td>.50</td>
<td>.50</td>
</tr>
<tr>
<td>3</td>
<td>1</td>
<td>9</td>
<td>.53</td>
<td>.47</td>
<td>.47</td>
</tr>
<tr>
<td>14&lt;sup&gt;b&lt;/sup&gt;</td>
<td>1</td>
<td>8</td>
<td>.47</td>
<td>.53</td>
<td>.53</td>
</tr>
<tr>
<td>15</td>
<td>1</td>
<td>8</td>
<td>.50</td>
<td>.50</td>
<td>.50</td>
</tr>
<tr>
<td>16</td>
<td>1</td>
<td>8</td>
<td>.53</td>
<td>.47</td>
<td>.47</td>
</tr>
<tr>
<td>17</td>
<td>2</td>
<td>8</td>
<td>.57</td>
<td>.43</td>
<td>.43</td>
</tr>
<tr>
<td>18</td>
<td>3</td>
<td>6</td>
<td>.50</td>
<td>.50</td>
<td>.50</td>
</tr>
<tr>
<td>19</td>
<td>3</td>
<td>5</td>
<td>.56</td>
<td>.44</td>
<td>.44</td>
</tr>
<tr>
<td>20</td>
<td>1</td>
<td>3</td>
<td>.50</td>
<td>.50</td>
<td>.50</td>
</tr>
<tr>
<td>21</td>
<td>2</td>
<td>3</td>
<td>.60</td>
<td>.40</td>
<td>.40</td>
</tr>
<tr>
<td>22</td>
<td>1</td>
<td>1</td>
<td>.33</td>
<td>.67</td>
<td>.67</td>
</tr>
<tr>
<td>23</td>
<td>2</td>
<td>1</td>
<td>.50</td>
<td>.50</td>
<td>.50</td>
</tr>
</tbody>
</table>

*Note. There was a total of 23 sessions. For Sessions 1 and 4–13, all participants had secondary teaching experience; thus, these sessions are absent from this table. After the cards in the rand-secondary bag had been exhausted, all subsequent participants drew from this stratum’s bag. These cards were in the rand-other bag. Participant 14 was the final participant that could have selected from the rand-secondary bag; however, because there was only one card left in that bag by the time Participant 14 entered the study, this individual would have automatically been assigned to the treatment group (with \( p = 1.00 \)). Therefore, prior to Participant 14’s arrival, the remaining rand-secondary card was added to this stratum’s bag to make the probability of being in either group close to .50.*
Conditional branching was used to avoid respondents answering questions that were not relevant to their experience. For example, if respondents answered Yes to Question 2, the survey software directed them to respond to Questions 3–6; if they answered No to Question 2, the software skipped Questions 3–6 and took them to Question 7. The Next buttons indicate page breaks in the questionnaire.

**Background Questionnaire**

*Welcome! The purpose of this questionnaire is to learn about your experience and knowledge in teaching, assessment, and vocabulary learning. These questions are relevant to this study because I will ask you to examine multiple-choice items on a vocabulary test and make a judgment about how many students (high school students in Hawaii) will be able to correctly respond to each item.*

I want to know if you have experience teaching because you might have an understanding of high-school students’ vocabulary abilities and of the way they will perform on each item on the test. You might also be familiar with the vocabulary abilities that are required among college students.

Similarly, if you have knowledge and experience in assessment, you might have an understanding of students’ thought processes when they read multiple-choice items.

And, if you have experience in the field of vocabulary and language learning, you might have an understanding of which vocabulary test items are easy and which are difficult.

*Directions*

This questionnaire will take about 10 minutes to complete. Please answer all the questions in one sitting. There are up to 22 questions, which are on several pages. Click the Next button to navigate through the pages.
1. Please enter the 4-digit ID code that you received (you can copy and paste the code below).

Part 1. Teaching experience

2. Do you have experience teaching middle- or high-school students?
   - Yes
   - No

Your experience teaching middle- or high-school students

You said that you have experience teaching students at the middle- or high-school level. The questions on this page ask more about this experience.

3. How many years have you taught at this level?
   - 1 year or less
   - between 1 and 3 years
   - between 3 and 5 years
   - 5 years or more

4. What grade level(s) have you taught?
   - 6th
   - 7th
   - 8th
   - 9th
   - 10th
   - 11th
   - 12th
   - other

5. Have you taught English or English language-arts to these students?
   - Yes
   - No

6. Have you taught middle- or high-school students in the state of Hawaii?
   - Yes
   - No

7. Do you have experience teaching college or university students?
   - Yes
   - No
Your experience teaching college or university students

You said that you have experience teaching students at the college or university level. The questions on this page ask more about this experience.

*8. How many years have you taught at this level?

- less than 1 year
- between 1 and 3 years
- between 3 and 5 years
- 5 years or more

*9. Were any of the courses (or other instructional activities) in education or in a field related to education?

- Yes
- No

*10. Were any of the courses (or other instructional activities) in assessment, testing, or measurement?

- Yes
- No

*11. Were any of these courses (or other instructional activities) in English language arts, applied linguistics, or a related field?

- Yes
- No

*12. Have you taught college or university students in the state of Hawaii?

- Yes
- No

*13. What subjects (or skill areas) have you taught to undergraduate students? (For example, biology, Hawaiian studies, anthropology, writing skills, math skills, etc.)
*14. Do you have experience teaching English or English language arts?
  
  ○ Yes
  ○ No

Your experience teaching English

You said that you have experience teaching English or English language arts. The questions on this page ask more about this experience.

*15. How many years have you taught in this subject area?

  ○ less than 1 year
  ○ between 1 and 3 years
  ○ between 3 and 5 years
  ○ 5 or more years

*16. At what level(s) have you taught?

  ○ middle school
  ○ high school
  ○ undergraduate level
  ○ graduate level
  ○ Other (please specify)

Part 2. Knowledge and experience in assessment

*17. You might have some experience in assessment. Please indicate how much experience, if any, you have in each of the following assessment activities.

<table>
<thead>
<tr>
<th>Activity</th>
<th>None</th>
<th>Some</th>
<th>A lot</th>
</tr>
</thead>
<tbody>
<tr>
<td>Developing classroom assessments</td>
<td>○</td>
<td>○</td>
<td>○</td>
</tr>
<tr>
<td>Developing large-scale educational assessments, such as for your state's department of education</td>
<td>○</td>
<td>○</td>
<td>○</td>
</tr>
<tr>
<td>Developing assessments for broad use, such as for a school, program, curriculum, or textbook</td>
<td>○</td>
<td>○</td>
<td>○</td>
</tr>
<tr>
<td>Writing multiple-choice test items</td>
<td>○</td>
<td>○</td>
<td>○</td>
</tr>
<tr>
<td>Analyzing test-takers' responses to the choices in multiple-choice items</td>
<td>○</td>
<td>○</td>
<td>○</td>
</tr>
<tr>
<td>Analyzing test statistics, such as item difficulty or item discrimination</td>
<td>○</td>
<td>○</td>
<td>○</td>
</tr>
<tr>
<td>Writing descriptors of content standards or benchmarks</td>
<td>○</td>
<td>○</td>
<td>○</td>
</tr>
<tr>
<td>Setting performance standards which determine cutscores on an assessment</td>
<td>○</td>
<td>○</td>
<td>○</td>
</tr>
</tbody>
</table>

*18. Have you taken any courses in assessment, testing, or measurement?

  ○ Yes
  ○ No
  ○ Can't recall
**Part 3. Your knowledge of theories of English language learning**

*19. You might have knowledge of theories of English language learning. How much knowledge do you have in each of the following subject areas?*

<table>
<thead>
<tr>
<th>Subject Area</th>
<th>None</th>
<th>Some</th>
<th>A lot</th>
</tr>
</thead>
<tbody>
<tr>
<td>Reading development</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Vocabulary development</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Academic English development</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Theories of language learning</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

*20. Do you consider English to be your native language?*

- [ ] Yes
- [ ] No

**For participants whose native language is not English**

*21. Do you believe you understand English well enough to easily read college-level textbooks without the aid of a dictionary?*

- [ ] Yes
- [ ] No
Is there something you'd like to add?

22. This is the final question.

As I mentioned on the first page of this questionnaire, I want to know if you have experience teaching because you might have an understanding of students’ vocabulary abilities and of the way they will perform on each item on a test you will be examining.

Similarly, if you have knowledge and experience in assessment, you might have an understanding of students’ thought processes when they read multiple-choice items.

And, if you have experience in the field of vocabulary and language learning, you might have an understanding of which vocabulary test items are easy and which are difficult.

If you think there is something in your experience which would be informative to me in this study, but which was not captured in the questions above, please share this information here. *(Leave blank if you have no additional information.)*

Thank you

Thanks for completing this questionnaire. See you soon!
REFERENCES

American Educational Research Association, American Psychological Association, &
Association.

Council on Education.

Testing, 26, 101–118. doi:10.1177/0265532209340194

have gone before!). Applied Measurement in Education, 9, 215–235.
doi:10.1207/s15324818ame0903_2

Brandon, P. R. (2004). Conclusions about frequently studied modified Angoff standard-
doi:10.1207/s15324818ame1701_4

L. Crocker & M. J. Zieky (Eds.), Joint conference on standard setting for large-scale
assessments (pp. 269–287). Washington, DC: NAGB–NCES.


cutting score procedures using generalizability theory. Applied Psychological


Hurtz, G. M., & Hertz, N. R. (1999). How many raters should be used for establishing
cutoff scores with the Angoff method? A generalizability theory study. *Educational

setting: Assessing “fit” to item characteristic curves. [Article]. *Applied Measurement
in Education, 22*, 120–143. doi:10.1080/08957340902754601

standard-setting judgments to cutoff scores on the item response theory θ scale.

of the assumptions in the Angoff standard setting method. *Journal of Educational


Kane, M. T. (1987). On the use of IRT models with judgmental standard setting
3984.1987.tb00284.x


Mahwah, NJ: Lawrence Erlbaum.


