# The Impact of Test-Taker Criticism on Educators' Question Review Self-Efficacy and Test Review Quality

BY

WILLIAM J. TRAPP
B.S., Western Illinois University, 1997
M.S., Northern Illinois University, 2007

DISSERTATION
Submitted as partial fulfillment of the requirements
for the degree of Doctor of Philosophy in Educational Psychology
in the Graduate College of the
University of Illinois at Chicago, 2018

Chicago, Illinois

Defense Committee:

Yue Yin, Chair and Advisor
Everett Smith, Jr.
Kathleen Sheridan
Michael Thomas
Theresa Thorkildsen
Alison Superfine, Mathematics, Statistics, and Computer Science

## ACKNOWLEDGMENTS

I would first like to thank my advisor and committee chair, Dr. Yue Yin, for her support, guidance, and perseverance. She provided honest and accurate feedback in a timely and professional manner. My research and writing quality greatly improved as a result of her guidance.

I would not have made it this far without the guidance from my former advisor, Dr. Carol Myford. She also preserved through my writing drafts and was a model of someone who transitioned gracefully from educational testing into academics.

I would also like to thank my cooperating teacher at my research site, Luke Wilcox. He worked with me patiently during the IRB approval process and went above and beyond in handling the logistics of the data collection events.

I would also like to thank the other members of my dissertation committee, Dr. Smith, Dr. Sheridan, Dr. Thomas, Dr. Thorkildsen, and Dr. Superfine for their expertise, insight, support, and helpful suggestions throughout my dissertation process. In addition, I would like to thank Dr. Stacey Horn, who provided excellent and helpful guidance during my dissertation proposal process.

Lastly, I would like to thank my family. My wife Kim, and my three children Holly, Jacob, and Lily supported me throughout my entire graduate education, and especially through the unique challenges that a dissertation provides.

## TABLE OF CONTENTS

## LIST OF TABLES

## LIST OF FIGURES

## LIST OF ABBREVIATIONS

| | |
|---|---|
| AERA | American Educational Research Association |
| ANOVA | Analysis of variance |
| ANCOVA | Analysis of covariance |
| APA | American Psychological Association |
| ETS | Educational Testing Service |
| IUA | Interpretative/use argument |
| NCME | National Council on Mathematics Education |
| SPR | Student produced response |
| 4MC | 4-option multiple choice |

## SUMMARY

Research has revealed the value of college admissions testing on the college application and admissions process. The combination of college admission test scores and the applicant's high school grade point average help the applicant and the college predict whether they will be a good fit together. Therefore, test stakeholders have high expectations of math question quality. Math question quality partially relies on the use of qualified question writers and multiple question reviews, including reviews by professional test developers and current math educators. Test takers are the largest group of test stakeholders and they are rarely consulted during the writing and reviewing of test questions. However, test taker question criticism may permit educators to confirm that they are finding and resolving all test taker concerns and may improve the quality of educator question reviews. Also, reviewing test taker criticism is a type of vicarious experience that may increase educator's confidence in their question review ability (also referred to as educator self-efficacy). The purpose of this study was to investigate the impact of test taker question criticism on educator self-efficacy and question review quality during the standardized test question review process.

In this study, first I recruited nine high school students as test takers. After giving them test review training, I asked them to review sixteen unused SAT math items. I further collected and summarized these students' test item criticism. Then I recruited sixteen educators and randomly assigned them into a control group or an experimental group. I facilitated separate question reviews: the control group of educators reviewed the questions without access to test taker criticism and the experimental group of educators reviewed the questions with access to test taker criticism. Participants self-reported their self-efficacy prior to and immediately after the question

review event. During each question review, participants identified question issues and, where possible, attempted to remove the issues with revisions to the question.

The results revealed both similarities and differences between the two groups of educators. Split-plot ANOVA of the pre- and posttest administrations of the appraisal inventory suggest that educator self-efficacy in both educator groups increased significantly from pretest to posttest. ANCOVA results suggest that the posttest scores of the experimental group were significantly higher than the posttest scores of the control group, after controlling for their pretest educator self-efficacy score. That is, the test taker question criticism seemed to help the experimental group increase their educator self-efficacy score more than the control group. In contrast, inclusion of test taker criticism in the question review process did not have a measurable impact on the question review quality. The question review audio transcriptions were coded and themes emerged were closely connected to individual statements and overall appraisal inventory concepts such as math content, construct-irrelevance, cognitive psychology, and consequences of testing. During the question reviews, the test takers did not successfully identify question issues but both groups of educators successfully identified multiple issues and resolved the issues by proposing changes to the questions. These findings have implications for test taker involvement in the question review process, as well as educator question review training and evaluation.

# I. INTRODUCTION

Due to the high stakes of college admissions testing, test stakeholders require high quality test questions. Test developers routinely invite educators who have rich content knowledge and teaching experience to review test questions. Educators can, to some degree, also represent test takers during question review. Test takers, as an important stake holder group, are not ordinarily invited to review test questions. Instead, stakeholders depend on educators to speak on behalf of test takers. Yet, without question criticism from test takers, educators may lack self-efficacy in their test question judgments. Such judgments made without confidence may result in superficial recommendations that reduce question quality rather than enhance it. I conducted this study to investigate whether exposure to test taker's criticism of SAT Math Test questions improved educators' self-efficacy in their ability to review test questions and improved the question review quality.

## A. Background

### 1. Test question review

In the United States, 88.2% of colleges describe college admissions test scores (SAT and ACT) as considerably or moderately important to first-time freshman admission decisions, and 72.7% describe the scores as considerably or moderately important to international freshman student admission decisions (Clinedinst, Koranteng, & Nicola, 2016). Therefore, test developers must ensure that college admission tests have high technical quality, as measured by indicators such as reliability and validity evidence. To ensure high technical quality, testing companies utilize multiple processes, such as a systematic review of test questions by educators. A high-quality test question review will result in the identification and resolution of undesired test question issues.

The evaluation of math questions on a college admissions test for undesired issues includes, but is not limited to the following question aspects: (a) whether the question has one correct answer, (b) whether the educators have taught or students have learned how to answer the question, (c) whether the educators have assigned or students solved similar questions in homework, and (d) whether certain question aspects interfere with the math skill or ability assessed. Several question aspects may interfere with the actual assessment of math skills and abilities. For example, visually-impaired test takers may be disadvantaged by certain text fonts and/or graphic presentations in a test question. Also, unfamiliar words or complicated sentence structures may disadvantage all test takers and are likely to disadvantage non-native English speakers more so. Particular words or phrases have different meanings in different parts of the United States and in different parts of the world, e.g., football means different sports in the US and the UK. Also, many words have multiple meanings. For example, the word "mark" may be used as a proper noun, a noun, or a verb. There are many ways in which a math test question may be assessing something besides mathematical skills and abilities. Therefore numerous stakeholders review questions from multiple perspectives to ensure that bias is reduced or eliminated. When multiple educators agree to remove questions with undesired issues and only approve questions that meet all expectations, then the result is high quality question review.

2. **Educator review and educator self-efficacy**

Recognizing the varying backgrounds and individual strengths of teachers, test developers of college admissions math tests routinely recruit a diverse set of math educators, including both high school teachers and college professors, in the test review process. Educators have the advantage of the requisite knowledge of a field. However, educator content knowledge and personal experiences may not compensate for an educator's lack of self-efficacy in their

question review ability. I have observed many educator question review events from the perspective of a test developer. Often, new educator participants are not confident in their question review ability. Lack of confidence likely stems from a lack of experience, but there are other ways to increase self-efficacy including, but not limited to, vicarious experiences. Test developers wish to report that all educator participants, of all backgrounds and experience levels, are confidently identifying and resolving issues with test questions.

Test question quality is one of several possible consequences related to educator self-efficacy. Teacher question review participants have an enormous opportunity to impact student success by permitting or denying specific test questions from appearing on a standardized test. Tschannen-Moran and Hoy (2001) suggest that teacher self-efficacy is related to the amount of effort invested in teaching and is also related to teachers' belief that they have an impact on student success or failure. If these relationships extend to educator question reviews, then educators with lower confidence in their question review ability may invest less effort in their question reviews and may not be attentive to their proposed recommendations on student success. Additionally, test stakeholders' confidence in a test score may weaken if it was determined that the educator participants did not believe that their decisions had any impact on the test scores that resulted.

By involving educators in the review process, test developers improve the test question quality and collect evidence to support the use of the test for college admissions. Also, test developers collect evidence to support the interpretation of the scores test takers receive. Lastly, because of the unique qualifications of educators, test developers utilize their comments to evaluate consequences, both intended (e.g. changes to curriculum) and unintended (e.g. changes in test preparation strategies), of the test use and test score interpretations.

3. **Test taker review**

Besides educators, another important group of test stakeholders is test takers. Test takers are the largest group of test stakeholders of college admissions tests. However, test takers do not participate in question reviews in the same way educators do. Several reasons prevent test developers from collecting similar data from test takers directly: First, if some test takers reviewed questions prior to taking the test while others did not have this opportunity, then the two groups of test scores would require different interpretations. Second, the logistics of organizing a review of secure materials by 17- and 18-year-old students from across the country is much more difficult than enlisting educators to complete the same task. Third, question review data collected from a small group of test takers will be less reliable and less generalizable than data collected from a small group of educators. Due to these barriers, educators commonly participate in the development process on behalf of test takers.

Although educators' expertise and proximity to test takers are unique and important, and can partially overcome the difficulties of involving test takers in the review process, test developers have little to no assurance that math educators reliably represent test taker perspectives when they evaluate math test questions. In addition, little to no research has investigated whether the resulting changes based on educator feedback would be rated as "improvements" by test takers. One exception is a study in which the target population was consulted during the development of a psychological assessment (Vogt, King, & King, 2004). The results of that study suggest that the participants enhanced the instrument as well as the overall validity evidence. It is unclear whether educators are appropriately representing test takers during the item review process, and I have been unable to locate research that has investigated the level of self-efficacy educators have in their ability to speak on behalf of test takers. Therefore, test developers may be unwise to assume

educator question review abilities, specifically the ability of educators to represent test takers in the question review process.

One possible solution to the issue of test taker representation is to expose educators to student criticism of math test questions in the form of a vicarious experience. Math educators already have the opportunity to ask students for their feedback regarding classroom assessment questions and discuss released questions from college admissions tests with their students. If exposure to student opinions of test questions is beneficial, then teachers will better understand student perspectives. Furthermore, if teachers have a stronger grasp on student perspectives, then the self-efficacy in their own judgments regarding questions for college admissions tests and their review quality may increase. Researchers have investigated the impact of test taker's opinions on educator question reviewers using qualitative research methods and new empirical research could provide more context and understanding of those findings.

## B. **Purpose of the Study**

In this study, I will examine whether exposure to student criticism of test questions can help improve educator reviewer's self-efficacy and the resulting question review quality.

To this aim, my research questions are as follows:

1. To what degree does educator exposure to test taker criticism of questions result in an increase in educator's self-efficacy regarding their question review judgments?
2. To what degree does educator exposure to test taker criticism of questions result in a higher quality question review during the question review process?

## C. **Approach of the Study**

I conducted an experimental study to answer both research questions. As test developers use quantitative and qualitative data from educator's question reviews as test validity evidence in

practice, I collected both quantitative and qualitative data during the study. Through the data analyses, I revealed how test taker criticism of test questions impacts the self-efficacy that educators have in their ability to review a variety of test question aspects. I also utilized the data to investigate impacts to question review quality.

## D. **Significance of the Study**

In this study, I compile educator question review expectations from a variety of sources in one location. This is a valuable resource for test developers and for researchers to reference in future research. Also, this study extends current understanding of teacher self-efficacy to educator confidence in their ability to review test questions. Lastly, the involvement of test stakeholders in the test development process is a frequent recommendation by critics, but the recommendation is rarely employed. This study implements one method of involving students in the question review process and how their involvement impacts the results of an educator question review.

# II. REVIEW OF LITERATURE

In this chapter, I first provide a brief overview of how educators review college admissions test questions. I then discuss why industry standards mandate an educator's question review as part of the test's validity evidence. Lastly, I examine the specific methods by which educators evaluate test questions and further suggest why educators' self-efficacy in their judgments may be increased if they are informed with test taker perspectives.

## A. Educator Question Review and Question Review Quality

Each year, educators from high schools, 2-year colleges, and 4-year colleges contribute to the development of college admissions tests by reviewing potential test questions. For the convenience of expression, I will simply refer to these reviewers as *educators* in the following discussion. Educators participate in the process because the stakeholders of these tests results – including the test takers and their educators, school administrator and counselors, and especially educators of college students – value their contribution. However, the contribution of educators may not be meeting the expectations of all stakeholders. In fact, some stakeholders will likely reject test results if they perceive that test question reviewers are ineffectively resolving issues with test questions and permitting flawed questions to continue in the question development process. Therefore, research into the review of test questions by educators is warranted to investigate whether test developers are meeting stakeholder expectations.

Educators review test questions for standardized tests using formal processes. For example, Delgado-Rico, Carretero-Dios, and Ruch (2012) described a systematic method for subject matter experts to participate in the development of the Spanish adaptation of the State-Trait Cheerfulness Inventory trait form. They considered factors such as the number of judges, judge qualifications, the specific judgments made, and the total numbers of ratings per participant. Although Delgado-

Rico et al. (2012) did not describe efforts to develop a college admissions test, the overall test development process is similar. In fact, corresponding methods have been described for nursing assessments (Berk, 1990; Grant & Davis, 1997) and for the development of an elementary school test in geology (Polin & Baker, 1979, April). As the test development method is nearly identical across disciplines, lessons learned in one discipline are often applicable to college admissions testing.

There are many observable characteristics of a high-quality question review. First, the participants of a high-quality question review must identify all issues present in the question and propose solutions to resolve the issues. Second, the participants must reach consensus on a single course of action for each question and all the participants should strongly agree with the course of action. Lastly, any changes to questions should not have introduced new issues. The results of a high-quality question review are high quality questions that multiple stakeholders approve for student administration. These questions are likely to be tried out and the subsequent statistical analysis should reveal appropriate difficulty and discrimination values. Also, high quality questions should not generate test takers irregularity reports following a test administration to indicate one or more issues with the question.

Critics have helped improve educator's question review process, and therefore the resulting question quality, by voicing their concerns in the literature. Sireci (1998) describes the aspects of an educator review that he encourages test developers to include. For example, if the group of test questions being reviewed does not adequately sample the content described, then educators should help uncover and highlight such inadequacies. Also, critics warn that educator ratings may be subject to confirmationist bias (Kane, 2006; K. E. Ryan, 2002; Sireci, 1998). Confirmationist bias may undermine educator judgments when teachers provide ratings they think test developers want

to hear instead of ratings that ensure integrity of the process and the resulting test materials, while critical feedback improves the resulting test validity evidence when they highlight weaknesses in existing processes and suggest possible solutions.

In addition to identifying weaknesses, fellow researchers scrutinize educator participation methods in an effort to hold test developers accountable to all test stakeholders. When educators review college admissions test questions, they provide a voice for many other stakeholders who are not directly involved in the process. Concerning college admissions testing, many thousands of educators are impacted by the decisions of the educator reviewers. Additionally, thousands of college admissions staff will receive and use test results to inform college admissions decisions. However, one often-overlooked aspect of educator's participation is that a very small number of teachers are participating on behalf of the largest group of stakeholders – test takers.

Educators are uniquely qualified to represent the voice of test takers when they review potential questions for college admissions tests for multiple reasons. First, educators have content expertise in the skills and abilities that test takers are expected to have. Second, educators have provided instruction to test takers and have formed opinions about the degree to which students can acquire the expected skills and abilities. Third, educators have administered their own tests to students and have had unique experiences in which they discovered how test takers of varying abilities and certain demographics responded to different questions. Lastly, educators have taken one or more college admissions tests and are aware of the consequences of test score interpretation and test score use. These unique qualifications provide a robust perspective from which educators may make judgments and recommendations about proposed test content. For the sake of all stakeholders, especially test takers, educators must participate in ensuring the quality of test questions and they must do so with some degree of self-efficacy.

## B. **Educator Self-Efficacy**

Self-efficacy is a person's belief in their ability to perform a specified task or achieve a specified goal (Bandura, 2006). In the context of social learning theory, a change in self-efficacy is not solely determined by one's environment nor by one's genetic dispositions (Bandura, 1986). Teacher self-efficacy is a specific facet of a person's self-efficacy. "A teacher's efficacy belief is a judgment of his or her capabilities to bring about desired outcomes of student engagement and learning, even among those students who may be difficult or unmotivated." (Tschannen-Moran & Hoy, 2001, p. 783). Evaluating test questions is a specific facet of teaching. Not all teachers have the opportunity to participate in a standardized test question review. For those who do, there is a learning curve during which they become acclimated to the process and the types of issues to identify and resolve. In this study I define educator question review self-efficacy (educator self-efficacy for short) as an educator's judgment of their capability of identifying question issues and determining question appropriateness.

Research about student achievement increasingly cannot ignore teacher self-efficacy as a variable. Serving as a bridge, teacher self-efficacy has relationships to both teacher variables (e.g. motivation, commitment, enthusiasm) and student variables (motivation, achievement, self-efficacy) (Muijs & Reynolds, 2015; Tschannen-Moran & Hoy, 2001). Those who study student achievement must attend to both teacher and student variables. However, researchers may be dissuaded from conducting such research due to the daunting number of possible teacher and student variables. Consequently, researchers may be passing up opportunities to study student achievement.

One possible solution that may encourage new research related to student achievement is to encourage researchers to utilize measures of self-efficacy. If an interesting relationship emerges

while using a self-efficacy measure, then the researcher may delve deeper into targeted areas such as motivation and enthusiasm. Self-efficacy research paved the way for teacher self-efficacy findings, just as teacher self-efficacy findings have paved the way for this study involving educator self-efficacy. Since little research exists concerning educator question reviews, educator self-efficacy is one of several approaches that researchers can use to make connections in new research related to student achievement.

Bandura (1997) suggested that changes to an individual's self-efficacy are linked to four primary sources: mastery experience, vicarious experience, social persuasions, and physiological and affective states. In a content review, Morris, Usher, and Chen (2016) categorized relevant research about teacher self-efficacy into which of the four primary source(s) were investigated. They found that mastery experience was a focus most often, followed vicarious experience and social persuasions. The questions on teacher self-efficacy instruments often attempt to collect information about one or more of these four primary sources. In the following paragraphs, I describe how each primary source influences the type of questions asked.

The primary source 'mastery experience' refers to perceptions of an individual's past teaching experiences. For example, a researcher may ask a teacher to rate their level of satisfaction or their level of success during the past year (Morris et al., 2016). In the context of test question review, a researcher may investigate the impact of mastery experience on educator self-efficacy by asking a teacher about their level of satisfaction or success on a previous question review event. Bandura (1997) cautions that questions about mastery experience should focus on teacher's perception(s) of the experience and not solely on the quantity of the experience.

The primary source 'vicarious experience' refers to perceptions from observations of other teachers and may include the consideration of oneself as a social model (Bandura, 1997).

Therefore, questions may ask teachers whether they value observation time or to rate the level of proficiency of an individual they observed. Concerning a test question review that they participated in, teachers may report on the value contributed to the conversation by other colleagues. Or, they may reflect on their performance or consider whether they provided a good model to other attendees.

The primary source 'social persuasions' refers to feedback from others. However, teachers may personalize the feedback more or less depending on their perception of the person's knowledge and/or credibility (Bandura, 1997). Morris et al. (2016) summarize several ways teacher self-efficacy may change due to social persuasions including, but not limited to: interpersonal support from other teachers, mentoring from other teachers, teaching feedback from other teachers, and feedback from students. In a study investigating the source of teacher self-efficacy change for an educator question review, a researcher may ask teachers about the 'messages' they have received about their question review performance; either from the test developer or from other teachers.

The fourth and final source 'physiological and affective states', refers to teacher "stress, fatigue, anxiety, and mood" (Morris et al., 2016, p. 4). Morris et al. (2016) also note that existing research focuses strongly on negative physiological and affective states and very little on positive states. Educator question review self-efficacy may increase or decrease depending on competing priorities, fatigue from the review or the combination of the review and teaching, as well as other issues related to physiological and affective states.

The relationship between teacher self-efficacy and teaching effectiveness continues to be a focus of research studies. As researchers understand this relationship better, their conclusions may help inform teacher education programs. There is a parallel, unexplored track of research regarding

teacher question review. The relationship between educator question review self-efficacy and question-review effectiveness is unexplored. When more is known about this relationship, test developers may use the conclusions to improve educator preparation and/or the question review process.

Educator self-efficacy may increase if educators participate in a vicarious experience in which they observe test takers critique test questions. In a pilot study, Trapp (2015) found that adding test taker's voice to the educator review process generated new and valuable information for the educators to consider. For example, the test takers had a variety of opinions about question difficulty, question wording, overall question appropriateness. Additionally, Trapp (2015) found that educators valued having the test taker comments and felt more confident in their own recommendations when they took these viewpoints into consideration. Initial findings suggest that the contribution of educators may be enhanced by the addition of test taker voice. To better understand this potential impact of test takers on educators, I reviewed the literature on including educator question review to support the interpretations of college admissions test scores.

## C. **Argument Based Validation: Including Educators**

Test developers have a large number of data sources from which they can derive test validity evidence and the test taker question responses are used for many validity analyses. With limited resources (people, time, money), test developers must choose which validity evidence to obtain from their available options. Validity evidence from an educator question review can be expensive due to travel costs, substitute teacher reimbursement, meeting space costs, and stipends. Each logistical decision about the educator question review process has an impact on the quality of the validity evidence obtained, so test developers must choose wisely.

Educator question reviews contribute to stakeholder support of the test use and test score inferences, and any validity evidence collected will permit stakeholders to increase or decrease their support. In this section I will describe the contemporary theory of test validity as well as the contemporary framework for test validation. Then, I will utilize the test validation framework to explicitly show how an educator review contributes to the quality of test questions, and therefore the validation of a college admissions test.

The standardized testing industry in the United States benefits from scholars who play an active role in developing the theory of test validity. As this theory has evolved, several books have proved invaluable to test developers, but none more so than *The Standards for Educational and Psychological Testing* (AERA, APA, & NCME, 2014), hereafter referred to as the *Standards*. The first edition of the *Standards* was published in 1966 and the most recent ($5^{th}$ edition) was published in 2014. While the *Standards* have no legal binding, the guidance provided within each edition is instrumental in helping test developers operationalize the contemporary theory of test validity.

As scholars have refined the meaning of *validity* over time, they have also refined guidelines for the process of test validation. The *Standards* defines validity as "…the degree to which evidence and theory support the interpretations of test scores for proposed uses of tests." (AERA et al., 2014, p. 11). Scholars emphasize that validity is not a property of tests, nor is it a label for one or more test score interpretations. Scholars have also shifted from writing about validity to writing about validation – the process of collecting and evaluating evidence to support the interpretations of test scores for proposed uses of tests.

The *Standards* is one of several guiding documents that mandate educator participation in the test development process. Standard 1.9 states that "When a validation rests in part on the opinions or decisions of expert judges, observers, or raters, procedures for selecting such experts

and for eliciting judgments or ratings should be fully described."(AERA et al., 2014, p. 25) Additionally, *Educational Measurement* (Brennan, 2006; Linn, 1989) contributes to the educator participation mandate. For example, Kane (2006) and Messick (1989) describe how educator participation was not an original mandate, then how stakeholders requested educator participation under the umbrella of content validity, and currently how educator participation is required under the umbrella of construct validity. Additionally, Schmeiser and Welch (2006) describe content-based judgments, Camilli (2006) and Zwick (2006) describe construct-irrelevant factors , and Zwick (2006) describes consequences of college admissions testing that educators may evaluate. The detail in these book chapters also explains why stakeholders require teachers to review test questions. Stakeholders not only require educator participation in the test development process, they require careful, systematic implementation of educator expertise.

Test developers document details regarding educator participation, along with all other validity evidence, in technical manuals. The same documents that mandate educator participation in the test development process also heavily influence other validity evidence that test developers include in technical manuals. Both the ACT Technical Manual (ACT, 2014) and the SAT Technical Manual (College Board, 2017) have validity chapters that draw heavily from the *Standards* and *Educational Measurement*. A common underpinning of both technical manuals is the contemporary theory of test validity described by Messick (1989).

Messick's unified framework helps test developers make connections between the theoretical underpinnings of a test that require evidence to support the test score interpretation(s) and use(s). However, Messick's framework does not specify a method to carry out and document such evidence. Building on Messick's theoretical framework, Kane (2006, 2016) has further developed and refined methodology to validate test score use and interpretation.

The first step described in Kane's methodology requires the test developer to state the test's purpose. Then, the test developer frames the interpretation/use argument, which Kane describes as "An *interpretation/use argument* (IUA) lays out the reasoning inherent in the proposed interpretations and uses of test scores, and thereby provides an explicit statement of what is being claimed" (Kane, 2016, p. 65).

During the last step of Kane's methodology, the test developer evaluates the coherence, reasonableness of inferences, and plausibility of assumptions of the IUA. Kane refers to this last step as the validity argument. In the next section, I use Messick's (1989) unified framework as the theory and Kane's (2006, 2016) validity argument as the methodology to describe potential test validation evidence from educator question review for a college admissions test.

1. **Test purpose**

There are similarities and differences between the purpose statements of the ACT and the SAT. Both technical manuals reference the use of the tests for college admissions decisions, course placement, and scholarship award decisions (ACT, 2014; College Board, 2017). The ACT technical manual goes on to reference additional uses by high school counselors, federal accountability testing, and determination for financial aid. With each test purpose listed, the test developer is informing stakeholders about the decisions that the test scores may inform.

2. **Interpretive/use argument**

The IUA is part of Kane's (2006, 2016) recommendation to keep test validation efforts directed and as concise as possible. Categories around which test developers create the IUA include: scoring inference, generalization inference, extrapolation inference, and decision inference. Kane placed these inferences in this order to highlight the "chain of inferences" (Kane, 2016, p. 66) followed from an observed test taker score to the decision made based on that score.

Based on the literature, I developed an IUA for college admissions testing, specifically for the educator review of test questions. In the following paragraphs, I describe each assumption within the four IUA categories.

*Scoring inference*. During the development of a college admissions test, many of the processes and procedures reflect an attempt to ensure that the responses provided by the test taker receive the correct score. The educator question review contributes evidence to support those efforts when test developers conduct the review in a way which addresses known assumptions. For example, researchers expect individual educator participants to have appropriate qualifications (Downing & Haladyna, 1997; Grant & Davis, 1997; Rubio, Berg-Weger, Tebb, Lee, & Rauch, 2003; Wynd & Schaefer, 2002). Educators have numerous qualifications of interest such as college degree level, college degree type, years of experience, number of publications, or other accomplishments relevant to the content of the assessment.

An additional assumption regarding scoring inference is that educator participants in question reviews will be effective contributors if they have proper training. For example, Polin and Baker (1979, April) ensured that the participants received training on the rating scale that they would use to evaluate questions. Prior to beginning a question review, educators typically receive an overview of the test development process, a summary of best practices for test questions, and an overview of why their participation is mandated.

Researchers also recommend that test developers use a systemic process to collect judgments from educator participants (Downing & Haladyna, 1997). The process should have clear directions for participants and, if the test developer uses an instrument to collect judgments, then the instrument must be appropriate for the type of test, e.g. norm referenced vs. criterion referenced (Rovinelli & Hambleton, 1976). Rovinelli and Hambleton also recommend that test

developers keep the judgments as simple as possible, and the number of ratings appropriate to prevent fatigue. Additionally, test developers should reduce undesired rater effects such as central tendency, leniency, and ratings made due to social desirability factors (D'Agostino, Karpinski, & Welsh, 2011; K. E. Ryan, 2002; Sireci, 1998). Lastly, this systematic process should attempt to avoid confirmationist bias.

The preponderance of the literature I reviewed described one or more possible educator judgments of individual test questions. These recommendations for question judgments may be summed up in a single assumption: Educators are providing judgments that aid test developers in the creation of high-quality questions, and therefore high-quality tests (Bridge, Musial, Frank, Roe, & Sawilowsky, 2003). This assumption seems obvious, but because educators play a limited role in test development, educators judge question aspects that they are best equipped to evaluate.

Four assumptions of scoring inference form a chain of support during validation: educator qualifications, training, question review process, and educator question judgments. Stakeholders use this particular chain of evidence to evaluate whether test taker responses are reliably receiving the correct score. The next chain of evidence is related to the generalizability of those test scores.

*Generalization inference*. Test developers use certain test development processes and procedures to ensure that groups of questions which make up a test are appropriate. Therefore, the assumption that educators receive proper training is both a scoring inference assumption and a generalization inference assumption. Specifically, Lennon (1956) suggests that educators need to receive enough training to conceive the universe of possible test questions. The educator training provides an overview that helps educators evaluate both question-level attributes and test-level attributes. This training often includes a reminder to the participants that they have unique

perspectives and should not assume that the other participants will see the items through the same lens.

An additional assumption regarding scoring inference is that groups of educators to have desired demographic characteristics (Downing & Haladyna, 1997; Grant & Davis, 1997; Rubio et al., 2003; Wynd & Schaefer, 2002). Target characteristics of the group of participants may include demographics such as gender, ethnicity, age, geographical region, type of educator (secondary or higher education), years of teaching experience, question review experience, or other. The makeup of the educator question reviewers is an important aspect because stakeholders link the diversity of the question reviewers to the diversity of question issues they are able to address. To address diversity expectations, test developers must also decide how many educators will participate (Rubio et al., 2003). Some researchers recommend a target number of educator participants to review a set of test questions, for example, generalizability theory may also be used to recommend this number (Crocker, 2003).

Researchers also expect that educator participants are able to predict, to a certain degree, test taker performance on test questions (Begeny, Eckert, Montarello, & Storie, 2008; Coladarci, 1986; Feinberg & Shapiro, 2009). Researchers expect the participants to distinguish between easy and difficult questions, as well as identify questions that discriminate between high-performing and low-performing test takers (J. J. Ryan, 1968). This evidence supports the generalization inference because groups of test questions must vary in difficulty to assess the desired construct.

*Extrapolation inference*. Educator question reviewers contribute to the extrapolation inference when they ensure that test questions do not discriminate against test takers on any other aspect aside from the intended test construct. Therefore, one of the assumptions of the extrapolation inference is that educator participants are capable of identifying construct irrelevance

in test questions including issues of opportunity to learn, question bias, and fairness (Bridge et al., 2003; T. Haladyna & Roid, 1981; Polin & Baker, 1979, April). Standard 3.0 (AERA et al., 2014) states that test development processes should minimize construct-irrelevant variance. Also, Standard 12.8 (AERA et al., 2014) states that high stakes decisions based on test scores should not include content which test takers have not had an opportunity to learn. Due to the nature of construct irrelevance, test developers utilize a diverse group of teachers to identify as many potential issues as possible because no single educator can speak for all aspects of construct irrelevance.

During my review of the literature, I unexpectedly did not locate researcher expectations regarding what test taker demographics should be represented during an educator question review. For example, stakeholders expect the College Board to ensure that questions are fair for English language learners (ELL). To do so, the College Board may include educator participants who work with ELL students and/or educator participants who themselves are ELL. Therefore, although not cited in the literature, I am including the assumption that educator participants can adequately represent historically disenfranchised test takers, such as ELL test takers, when they review test questions.

I have defined just two assumptions related to the extrapolation inference, but by no means do I intend to imply that this inference is less important than any of the other inferences. If the validity evidence to support these two assumptions is not adequate, stakeholders may not reject the test the same as they would for any of the other inferences, including the decision inference.

*Decision inference*. All assumptions regarding decision inference are based on the expectation that educator participants can anticipate whether a test question will contribute towards unintended consequences (K. E. Ryan, 2002). Examples of unintended consequences include

inappropriate test preparation, narrowing of the instruction provided by teachers, or inappropriate uses or interpretation of test scores. This category is broad, and training educators in such issues is not an easy task. When educators identify issues related to consequences, their comments are often anecdotal, unprompted and are stated similarly to "If I were teaching [math course] and I saw this question on a released test for [test name], then I would [consequence to instruction].” The *Standards* stress how multiple stakeholders, including the test developer and educator participants, must help bring attention to both intended and unintended consequences of testing.

Based on the literature reviewed, I drafted an IUA for educator question review. The result is a robust set of expectations. The resulting IUA is shown in Table I and includes 12 separate assumptions.

**Table I**

IUA AND EDUCATOR ITEM REVIEW ASSUMPTIONS

| Inference | Assumption for Educator Item Review Evidence |
|---|---|
| Scoring | 1. Individual educator participants have appropriate qualifications<br>2. Educator participants receive proper training in the specific question review tasks completed.<br>3. Test developers to use a systemic process to collect judgments from educator participants<br>4. Educators are providing judgments which aid test developers in the creation of high quality questions, and therefore high-quality tests. |
| Generalization | 1. Educator participants receive proper training concerning the test construct to the point they may conceive the universe of possible test questions.<br>2. Any and all groups of educator participants have desired characteristics such that the recommendations put forth are not unique to those assembled.<br>3. Educator participants are able to predict, to a certain degree, test taker performance on test questions. |
| Extrapolation | 1. Educator participants are capable of identifying construct irrelevance in test questions including issues of opportunity to learn, question bias, and fairness<br>2. Any and all groups of educator participants can adequately represent historically disenfranchised test takers when they review test questions. |
| Decision | 1. Educator participants can evaluate whether a test question will contribute towards unintended test takers consequences such as anxiety, de-motivation, and/or fatigue.<br>2. Educator participants can evaluate whether a test question will contribute towards unintended instructional consequences such as a narrowing of the curriculum, inappropriate test taker strategies, etc…<br>3. Educator participants can evaluate whether a test question will contribute towards unintended test use consequences such as misinterpretations and/or misuses of test scores. |

In summary, I developed an IUA using the underlying assumptions of an educator question review for college admissions testing. In some cases, the literature states the assumptions while in others the assumptions are not explicitly stated. Many of the assumptions begin with, "Educator participants can…", language that reveals how critical educator self-efficacy is to test validation. One of the many benefits of a well-developed IUA is helping direct test validation efforts and

making those efforts as concise as possible. In this particular case, the IUA is essential for generating an appraisal inventory for evaluating educator self-efficacy.

3. **Validity argument**

The validity argument is a body of evidence that the test developer has logically organized to evaluate the inferences and assumptions of the IUA. Each aspect of test development requires extensive preparation and appropriate methodology, including educator question review.

Test developers consider several limiting factors when preparing for an educator question review. The factors can be separated into two categories: logistical and process. Logistical factors include test developer staff time, external reviewer time, costs related to conducting the review, and the scope of the review. Process factors include when a question is reviewed, who reviews and how they are trained, judgments solicited, reconciling judgments, summarizing judgments and reporting results. For the purposes of this study, I do not discuss logistical factors any further except to say that quantity of evidence is not necessarily preferred over quality of evidence. The same could be said regarding process factors.

There are multiple points during the test development process at which educators may review questions. A chronological list of test development components is as follows, "overall plan, domain definition and claims statements, content specifications, item development, test design and assembly, test production, test administration, scoring, cut scores, test score reports, test security, and test documentation" (Lane, Raymond, Haladyna, & Downing, 2016, p. 4). Educators may review questions at one or more of these time components, but usually do so either prior to test design and assembly, after a small-scale item tryout, and/or after test administration (with data). The assumptions of the IUA do not mandate more than one question review. However, test developers may also ask educators to review specific groupings of items including complete test

forms. Therefore, test developers may choose to do one or more reviews depending on their available resources.

Test developers may document selected demographics and/or qualifications of educator participants in the technical manual, such as, a summary of participant educational background and teaching experience including courses taught and number of years of experience. In the absence of summarized demographic data, the test developer may choose to list the minimum criteria for individual educators and/or groups of educators.

Test developers may also document the training provided to educators in advance of a question review. Details about the training may include a summary of the content covered, the length of the training, and any tasks that the participants must complete successfully to continue participating. Documenting such information in the technical manual should be done in a way to support the assumptions of the IUA.

The bulk of the literature discusses the educator judgments solicited and methods to summarize the judgments obtained. In the next section, I review these aspects of educator question review along with how test developers document the validity evidence in technical reports.

## D. Educator Contributions

Educators make both direct and indirect contributions when they review test questions. Direct contributions include judgments for individual test questions as well as any adjustments to their ratings while reaching consensus. Indirect contributions include analyses that the test developer completes following the submission of educator judgments.

### 1. Direct contributions

Researchers provide several recommendations for eliciting educator judgments. If test developers receive educator judgments prior to question try-out, they are able to revise the

questions to address the educator concerns. However, if test developers receive educator judgments after question try-out, then judgments are typically limited to 'approved as is for a test form' or 'rejected and never used on a test form.' Either way, educator judgments provide valuable data for test developers.

The most frequently cited educator judgment is a test question's match to the description of the content assessed on the test. Several rating formats exist to elicit a test question's match to the content assessed including Dichotomous rating (Polin & Baker, 1979, April; Schmeiser & Welch, 2006), Likert format (Rovinelli & Hambleton, 1976; Rubio et al., 2003; J. J. Ryan, 1968), Matching (D'Agostino et al., 2011; Herman, Webb, & Zuniga, 2003; Li & Sireci, 2013; Rovinelli & Hambleton, 1976), and Item pair comparison (Sireci & Geisinger, 1992, 1995).

Each rating format requires different directions and the resulting ratings may be analyzed and reported differently. In addition to content match, many of these rating formats are also appropriate for the following educator judgments. In the following paragraphs, I describe ten additional judgments that test developers may ask educators to consider while reviewing test questions.

*The question matches the intended cognitive alignment*. For example, Bloom's taxonomy (Bloom, Englehart, Furst, Hill, & Krathwohl, 1956), Marzano's hierarchy (Marzano, 2001), Webb's depth of knowledge (Webb, 1997), or cognitive rigor matrix (Hess, Jones, Carlock, & Walkup, 2009) are cognitive alignment organizers. Test developers use a cognitive alignment to report additional detail about how a student is interacting with the content assessed. For example, the test developer may ask a question that requires a memorized response, or they may ask a non-routine question about the same content.

*The question is relevant to the test construct.* (Delgado-Rico et al., 2012; Polin & Baker, 1979, April; Rovinelli & Hambleton, 1976; J. J. Ryan, 1968; Sireci, 1998; Yaghmale, 2009). Ratings require participants to indicate how well the question being asked is assessing the expected content assessed. For example, a question may assess whether students can add whole numbers, but a question that requires students to add ten different whole numbers may be rated less relevant than a question that requires students to add two different whole numbers.

*The question uses language to strive for clarity.* (Bridge et al., 2003; Delgado-Rico et al., 2012; Grant & Davis, 1997; Rubio et al., 2003; Wynd & Schaefer, 2002; Yaghmale, 2009). Ratings ask participants to indicate whether a test question utilizes language correctly and concisely to convey the information

*The question stem uses language to reduce or eliminate ambiguity.* (Delgado-Rico et al., 2012; Yaghmale, 2009) This criterion requires an evaluation of whether test takers can interpret a test question in multiple ways. For example, a test question about a sweater on sale may reference the price of the sweater while the answer may be different depending on whether that price is interpreted as the original price or the price with the discount applied.

*The question is technically sound and defendable from criticism*. (Delgado-Rico et al., 2012; T. Haladyna & Roid, 1981; Polin & Baker, 1979, April; Schmeiser & Welch, 2006; Sireci & Geisinger, 1995). College admissions tests are high stakes so every point counts. For instance, educators may be asked whether, in their expert opinion, a test question is impervious to a challenge from a test taker.

*The relationship between the question and the possible answers ensures reliable scoring of correct and incorrect responses.* (Polin & Baker, 1979, April; Schmeiser & Welch, 2006). Test directions, along with a clearly worded question stem, may not properly distinguish why one

possible answer receives credit and another does not. Test takers can be creative when they are answering a question, therefore the connection between the question wording must help the test taker understand what response(s) will and will not receive credit.

*The question conforms to specified organization, style, and formatting rules*. (Polin & Baker, 1979, April; Schmeiser & Welch, 2006). Each test typically has its own rules for style, grammar, and formatting, and educators can help identify when a question is inconsistent with the rules.

*The question is appropriately presented to conform with the test directions.* (Polin & Baker, 1979, April). Each test has its own directions, and educators can help identify when a question is not aligned with a given rule or might be misinterpreted based on the directions.

*The question difficulty is appropriate, based on an estimate of the percent of test takers who will answer the question correctly*. (Coladarci, 1986; J. J. Ryan, 1968). J. J. Ryan (1968) asked 59 high school teachers to estimate item difficulty and summarized, "a fair proportion of teachers were able to provide statistically reliable estimates of the actual item difficulties" (p. 305).

*The question discrimination is appropriate, based on an estimate of whether a greater number of proficient test takers will answer the question correctly than will less proficient test takers*. (J. J. Ryan, 1968). Test developers strive to create items at a variety of difficulty levels, and for the percent of high-ability students who answer the question correctly to be higher than the percent of low-ability students who answer the question correctly. This judgment is not intuitive. Sometimes, unintended clues can assist less proficient students. For example, a key word in the question may be repeated in the correct answer of a multiple-choice question, but not repeated in the other answer choices. Students who can utilize such clues are often referred to as test-wise.

Educators are the best stakeholders to provide these content-based judgments due to their content expertise and classroom experience. For this reason, the same ratings from school administrators or school counselors do not help convince stakeholders that the test is appropriate for its intended use. In addition to content-based judgments, researchers also suggest that educators evaluate whether construct irrelevant issues are reducing question quality.

Educator judgments related to construct irrelevant factors identify issues that may alter performance for one or more subgroups of test takers, resulting in test scores that do not reflect their ability in terms of the construct measured. These mitigating factors may inflate the scores of some test takers and/or decrease the scores of others. Haertel and Lorie (2004) describe several examples of mitigating factors: test taker motivation, test taker anxiety, item format familiarity, text complexity (including vocabulary, sentence structure, sentence length, and other grammatical factors), and background knowledge and/or shared experiences.

Other sources of construct irrelevance are classified under the umbrella term of "fairness." The *Standards* devote an entire chapter to test fairness. Fairness is not well defined in the testing industry because different stakeholders use the term in different ways. However, resources such as the *Standards* and the *ETS International Principles for Fairness Review of Assessments* (Educational Testing Service, 2009) provide guidance for reviewing test questions from a fairness perspective. Additionally, researchers provide guidance on selecting and training participants for reviewing items through a fairness lens (Camilli, 2006; Zieky, 2016).

An additional umbrella term used to summarize selected construct irrelevant issues is universal design (Johnstone, Altman, Thurlow, & Moore, 2006). Universal design principles are adapted from architecture and provide guidance on making test questions accessible to test takers according to their content expertise. Just as certain properties of doorways facilitate the entry and

exit of a person in a wheel chair better than others, certain properties of tests facilitate the test administration to test takers. For example, adjustments to text font and size, as well as line spacing and column widths, may reduce reading difficulties for certain test takers. Educators who are well versed in universal design are especially helpful with identifying problematic graphics that will not display appropriately when produced in a large print format and/or not translate well to a Braille format.

Scholars have also investigated construct irrelevant factors related to cognitive psychology. Question development and review from this perspective involves labeling questions with traits in an effort to track the mental effort that test takers are exerting (Mislevy, 2006). Two such traits are working memory and long term memory (Pellegrino, Chudowsky, & Glaser, 2001). Working memory concerns the information that test takers must hold in their minds while answering a question. Long- term memory concerns the information that students bring with them and must access to answer the question. Considering such question traits can help test developers and educator question reviewers in eliminating questions that are not appropriately assessing the test construct.

In addition to content, construct-irrelevant, and cognitive psychology judgments, educators may also judge aspects related to unintended consequences of test score use or test score interpretation. For example, college admissions testing may cause anxiety for some test takers. Certain test question aspects may exacerbate test taker anxiety, causing test takers to sleep poorly the night before the test, utilize inappropriate test preparation, or other consequences that the test developer has not foreseen. In addition to anxiety, T. M. Haladyna and Downing (2004) connect motivation-related issues to potential unintended consequences. Test taker motivation may impact test preparation activities and test-day performance. In some cases, test takers are aware of a

certain test score use and interpretation while in other cases, they do not understand the consequences of their performance (K. E. Ryan, Ryan, Arbuthnot, & Samuels, 2007). One last example of unintended consequences is the influence of a test on curriculum and/or instruction. Test content assessed, or not assessed, may result in how teachers prepare students, including what content is taught, how it is taught, and when it is taught. Because of the ambiguity and the vastness of possible unintended consequences, as well as factors outside of educator control, educators are not the only stakeholder group that can provide guidance to test developers in this area.

Educators may provide a variety of judgments related to test questions; due to logistical constraints, test developers must choose how many times to collect judgments and which judgments best support the validity argument. According to Kane (2016), test developers need not collect every possible type of judgment. But it is in the test developer's best interests to collect the strongest possible evidence to support the IUA. Therefore, educator self-efficacy may be an important consideration as test developers decide which content-based judgments, construct-irrelevant judgments, and judgments regarding consequences to request from educators. For example, if educators self-report different levels of self-efficacy regarding their ability to judge test question alignment and test question difficulty, then test developers may revise their training. Also, if educators report low self-efficacy for one or more test question criteria, then test developers may seek either better prepared educators or seek groups of educators who can balance each other's strengths and weaknesses. In addition to the direct evidence provided by educators, test developers can also summarize those judgments to further support the validity argument.

2. **Indirect contributions**

In addition to the educator judgments described in the previous section, researchers also describe how the judgments may be analyzed for additional validity evidence. These analyses

may be classified as either reconciling disagreements or summarizing judgments. In this section, I summarize the analyses that test developers can perform on educator judgments.

When a group of educators review the same group of test questions, the educators will likely disagree on one or more judgments so the test developer will need to determine how to resolve these differences. Test developers can accept and reconcile ratings on their own or seek consensus with the educators. Fink, Kosecoff, Chassin, and Brook (1984) proposed a number of consensus methods and, as in the case of other educator judgments, the method to solicit consensus will be based on the logistical constraints of time, cost, and scope of work. Test developers can document the reconciliation process and results as part of the validity argument to support the IUA.

The test developer's choice of educator judgment(s) will determine the types of analyses they may conduct on the resulting data, such as: (a) Descriptive statistics of the judgments (Herman et al., 2003; Webb, 1997). The test developer may report the number of ratings made, the mean rating, or other statistic(s). (b) Interrater reliability (Grant & Davis, 1997; Herman et al., 2003; Polin & Baker, 1979, April; Rovinelli & Hambleton, 1976; Rubio et al., 2003). Educators do not always agree and when test developers summarize their ratings, the differences may become hidden. Reporting interrater reliability allows stakeholders insight into the level of agreement between the participants. (c) Validity index (Aiken, 1980, 1985). Aiken describes how test developers may collect educator test question ratings on an ordinal rating format (e.g. low, medium, high) and use the validity index to report whether or not the results are statistically significant. (d) Comparison of educator difficulty prediction to actual question difficulty (Begeny et al., 2008; Coladarci, 1986; Eckert, Dunn, Codding, Begeny, & Kleinmann, 2006; Feinberg & Shapiro, 2009; Hoge & Coladarci, 1989). When educators estimate question difficulty, the test

developer may calculate the actual item difficulty after the test administration and compare the predicted to the actual value. (e) Multidimensional scaling (D'Agostino et al., 2011; Hopkins, George, & Williams, 1985; Li & Sireci, 2013; Sireci, 1998; Sireci & Geisinger, 1995). This methodology allows test developers to visually investigate relationships between multiple properties of test questions. For example, Li and Sireci (2013) modeled how multidimensional scaling may be used to investigate the relationships between the content and cognitive alignments of multiple items and multiple raters.

Test developers are wise to select the indirect educator contributions that best support the IUA, and are not required to conduct every possible analysis. They have many methods, direct and indirect, to gather construct validity evidence resulting from educator question review judgments. These educator judgments are collected in context as part of an intentional process, which is embedded within the overall test development process. These processes are mandated by the *Standards* to reduce the misuse and misinterpretations of test scores. Under ideal circumstances, educator self-efficacy is adequate so that the largest group of stakeholders – test takers – are protected from test score misuse and test score misinterpretation. Test takers have little to no voice in the test development process, therefore the educator question review contribution is an opportunity for them to be spoken for.

## E. **Indirect Contributions from Test Takers**

Test developers are not required by industry standards to conduct a question review with test takers. However, test developers have utilized test takers to obtain information about test taker quality. I refer to such test taker contributions as "indirect" because the test taker criticism will not be used to evaluate the quality of the questions under review. Rather, the questions under review

are control variables and test developers are evaluating the test taker perspectives. There are multiple examples in the literature where researchers involve test takers indirectly.

In some studies, researchers are interested in how test takers interact with the test content and they are not interested in evaluating test question quality. For example, Katz, Bennett, and Berger (2000) investigated the impact of item type on how test takers answered test questions. Katz et al. (2000) asked the high school study participants to think aloud as they solved questions on the SAT math test. The study revealed that the test takers did not always use the solution strategy predicted by the test developer. Also, the results of the study suggest that solution strategies are not dependent on item type. Another study involving indirect test taker participation revealed that the cognitive alignment of a test question did not reflect how all test takers solved the same problem (Gierl, 1997). Gierl asked middle school test takers to think aloud as they solved test questions for a math test. Based on students' cognitive processing, Gierl classified their thinking using Bloom's taxonomy (Bloom et al., 1956). The results of the study indicated that "the cognitive processes expected by item writers matched the processes used by students in only 54% of the cases" (Gierl, 1997, p. 30). A final example of test taker contribution to content-based judgments of test questions involved The National Education Longitudinal Study of 1988. Hamilton, Nussbaum, and Snow (1997) conducted a factor analysis of the national data set and identified three dimensions. They then asked both middle- and high-school students to think aloud when solving selected test questions with the intent to strengthen the interpretation of test taker scores. Some of the findings included when test takers learned particular content, how knowledge of certain terms impacted performance, and how strategies of successful students differed from those of unsuccessful students. This study is yet another example of how indirect contributions from test takers will benefit educators, possibly increasing their self-efficacy, as they judge certain

aspects of test questions. In each of these examples, the data is more useful in understanding how test takers respond to questions than how test questions may be improved.

In addition to content-related variables, researchers also use data from test takers to investigate construct-irrelevance. For example, Gallagher and De Lisi (1994) investigated why selected items on the SAT math portion were answered correctly by differing percentages of male and female test takers. The researchers utilized existing data to identify the problematic questions. Then, Gallagher and De Lisi used structured interviews of high school test takers to investigate possible connections. Prior to this study, test developers and educators were unaware that males utilized certain solution strategies more often than females and vice versa. In fact, it is possible that educators are unaware of this phenomenon in their classrooms. I suspect that test takers have more to teach both test developers and educators about this and other types of construct irrelevant factors.

Researchers frequently cite language ability as a construct irrelevant factor in tests of math ability. In some cases, the language of the test may be a factor for non-native speakers, while in others, the language may be problematic for those whose first language is English. Serder and Jakobsson (2014) investigated language on the Programme for International Assessment and Leighton and Gokiert (2005) investigated language on the School Achievement Indicators Program Science Assessment. Both researchers investigated various aspects of language and utilized different methods to elicit comments from the test takers. However, the results of each study identified aspects of the language in test questions that are not accomplishing what the test developer intended. For example, (Serder & Jakobsson, 2014) concluded that test developers and test takers had very different conceptions of the meaning of "everyday life" contexts in science questions. A common high school math question on a standardized test provides a context and

requires the test taker to identify the variable, assign a letter to represent the variable, and then create an algebraic equation using that letter to represent the context. Test developers and educators strive to use contexts that are accessible to all test takers so that the probability that they answer the question correctly is based strongly on their mathematics ability and very little on their reading comprehension ability. It seems that test takers have much to teach test developers and educators about how accessible these contextual questions are.

The literature contains many more studies regarding the content and construct-irrelevance of test questions than the consequences of testing. In a research study concerning an agricultural assessment, Gulikers, Biemans, and Mulder (2009) asked multiple stakeholder groups to evaluate an existing test form. Regarding consequences of testing, two groups of test takers, a group with limited coursework and a group who completed advanced classes, provided judgments on whether the agricultural test should contribute positively to a test taker's motivation to pursue further education in agriculture. On a 5-point rating scale, only the test taker group who completed advanced classes had motivation ratings significantly above 3. Therefore, the authors concluded that the agricultural test is struggling to meet the motivation criterion. This study is the only one I located in which students are asked to evaluate test content on the basis of consequences of testing.

All three categories of direct educator judgments in the literature (content-based, construct-irrelevance, and consequences of testing) are present in the literature as indirect judgments provided by test takers. However, there are several limitations to the usefulness of these findings for college admissions testing. First, the test purposes in these examples are different from the test purposes of college admissions testing. Also, none of the data collection methods included a training in which test takers developed skills to evaluate test questions. Had these research students resulted in conclusions that generalized to college admissions tests, then test developers could

theoretically use those conclusions to train educators. Due to the absence of any data about the relationship between test taker voice and educator self-efficacy along with the lack of generalizability with existing studies, direct contributions from test takers need to be investigated first.

## F. Direct Contributions from Test Takers

Unfortunately, there are very few research studies in which test takers provide direct judgments concerning the quality of test questions. In fact, the only two research studies I was able to locate concerned tests in the medical field. Stewart, Lynn, and Mishel (2005) and Schilling et al. (2007) elicited judgments about questions from children under the age of 18. These "test-takers" were utilized as "experiential experts," or persons whose medical histories make them uniquely qualified to participate in the question review process. Both studies concluded that experiential experts played a valuable role in the process and Schilling et al. (2007) also emphasized that including test takers was an expression of our research ethics in terms of "participation, inclusion, and collaboration" (p. 365).

Due to the absence of research studies in which test takers evaluate test questions, I conducted a pilot study (Trapp, 2015). Test takers reviewed questions similar to those found on the math portion of the SAT and provided feedback. The resulting coding scheme classified their comments under the themes alignment, wording/clarification, or formatting. Therefore, the test taker participants contributed both content-based and construct-irrelevance feedback. Following the test taker judgments, educators reviewed the same questions – first without the test taker comments and then again with these comments. I concluded that, as did Stewart et al. (2005) and Schilling et al. (2007), experiential experts play a valuable role when they participate in the process.

Given the value that test takers add by both direct and indirect participation, I hypothesize that test taker contributions can enhance the judgments provided by educators. The *Standards* do not mandate taker participation in the test development process in the same way they mandate educator participation. Therefore, until the benefits of direct participation by test takers can be demonstrated, efforts should be directed at enhancing educator judgments.

I have found test takers' contribution to the educator review process (Trapp, 2015). However, the pilot study did not use experimental design, hence a cause and effect relationship could not be determined. Also, the pilot study revealed that both the student and educator test rating forms need to be clarified. Finally educators' self-efficacy was not measured in the pilot study. Therefore, this proposed study seeks to improve upon the pilot study by using an experimental design, improved test rating forms, and added self-efficacy as an outcome variable.

Test takers are a historically silent majority in the test development process. This study is a departure from indirect test taker involvement. Eight is not a large number of student participants, but direct involvement should be encouraging to test takers. Additionally, the test takers who participated in this study did so using the same process as the educator participants. By involving them fully in the process, and ensuring the process was focused on the resulting quality of the test questions developed, it is my desire that test takers feel well represented and heard.

## G. **Summary**

In this chapter, I provided an overview of the contribution that educators make to test question quality when they participate in the development of a college admissions test. Test developers must choose how and when educators participate in the process; the method selected has implications for time, cost, and the amount of work that can be accomplished. I have also described how educators speak for test takers when they provide judgments regarding question

content, construct-irrelevance, and/or consequences of testing. The creation of an IUA, which a test developer investigates through a validity argument, helps narrow down which evidence is required to support any test use(s) or test score interpretation(s). Lastly, I have hypothesized that the validity argument may be strengthened by direct involvement by test takers in the question review process. In the next chapters, I report the results of experimental study that I conducted to investigate the impact of test taker's criticism on educators' self-efficacy of test review and the resulting test question quality.

In particular, I investigated and answered the following research questions:

1. To what degree does educator exposure to test taker criticism of questions result in an increase in educator's self-efficacy regarding their question review judgments?
2. To what degree does educator exposure to test taker criticism of questions result in a higher quality question review during the question review process?

This study is the first that utilizes student feedback in the question review process. Also, this study is the first to investigate the impact of test taker criticism on educator self-efficacy and the quality of the test question review.

# III. METHODS

## A. Introduction

Test takers are the largest group of stakeholders regarding college admissions testing, and educators participate in the test development process on behalf of test takers. Unfortunately, little to no research is available concerning how well educators are representing test takers during the test development process. To fill this gap, I examined whether the vicarious experience of exposure to test taker criticism of SAT Math test questions results in higher educator self-efficacy and higher test question quality. Although other primary sources may influence educator self-efficacy including mastery experience, social persuasions, and physiological and affective states, this study focuses on a common mastery experience for the control and experimental groups and a single vicarious experience for the experimental group only. This chapter describes the method for the study.

Due to the involvement of human subjects, I applied for and received approval from the Institutional Review Board (IRB) at the University of Illinois at Chicago. Initial contact with potential participants included a summary of their involvement and either the Student Consent form or the Educator Consent Form (see Appendix A). Lastly, I obtained a signed consent form from each participant before proceeding.

## B. Research Design: Question 1

To answer research question 1, I conducted an experimental study and utilized a pretest-posttest control group design (Kerlinger & Lee, 2000). I randomly selected educator participants from the population and then randomly assigned them to either the experimental group or the control group. I utilized the appraisal inventory that is described later in this chapter as the pretest

and posttest instrument. The experimental manipulation, exposure to test taker criticism, was administered to the experimental group and withheld from the control group.

Using the pretest-posttest control group design, I examined whether and to what degree the exposure to test taker criticism of test questions impacted educators' self-efficacy. In particular, I measured educator self-efficacy with the appraisal inventory at two times and used the initial ratings as the covariate.

## C. Research Design: Question 2

I employed a mixed method research design to answer research question 2 using both quantitative data (e.g., appraisal inventory responses) and qualitative data (e.g., the audio recorded discussions). The quantitative data provides committed responses to question judgments and the qualitative data provides an opportunity to persuade others or hedge bets. As the relationship under investigation was not previously studied, I did not assign priority to either the qualitative data or the quantitative data so I remained flexible to emerging results. Both types of data were collected simultaneously. After data collection, I analyzed these data together. Creswell (2002) refers to this method as a triangulation design.

## D. Setting

I collected all data at the same location at a high school in Michigan. All three data collection events were held in the media center at the cooperating high school. The room was set up with multiple round tables that were arranged so that all participants could view the projected screen, as well as the other participants. There were multiple desktop computers in the center of the round tables.

## E. **Participants**

Due to time and resources available, I conducted this research using student and educator participants from a high school in Michigan, along with post-secondary educators at nearby colleges. In total, nine students, eight secondary school educators, and seven post-secondary educators participated in the study.

Upon receipt of approval from the IRB and approval from the cooperating school, I began recruiting student participants. First, I worked with a cooperating teacher and a school counselor to compile a list of eligible students. To be eligible, students must have taken the SAT within the past 15 months and be adults at the time of data collection. From the list of eligible students, I randomly selected eight participants. The cooperating teacher and the school counselor made initial contact with the students. When students declined to participate, I randomly selected replacements and provided the names to the cooperating teacher and school counselor. When students agreed to participate, the cooperating teacher or the school counselor obtained a signed consent form from the participant (see Appendix A). Multiple students indicated their desire to participate but would not confirm their availability to attend. Therefore, I recruited additional students to ensure that a minimum of eight participants were in attendance.

On the day of the data collection event, ten students arrived. However, one of the students could not stay for the entire event and the student's ratings and responses are not included in the data analysis or discussion. Table II shows the school's race and ethnicity information for the 2014-2015 school year along with the race and ethnicity information for the nine student participants. Overall the student participants have diverse ethnicity background, which is similar to their school. However, more female students than male students participated in the study. I

originally planned to recruit students with different levels of SAT scores. But I could not obtain IRB approval to use that information.

**Table II**

DEMOGRAPHIC SUMMARY OF STUDENT PARTICIPANTS

| | Percentage of Students in Cooperating School | Percentage of Student Participants |
|---|---|---|
| Primary Race/ Ethnicity | | |
| American Indian/ Alaskan | 0.3% | 0.0% |
| Asian/ Pacific Islander | 15.0% | 22.2% |
| Black | 30.1% | 11.1% |
| Hispanic | 11.3% | 33.3% |
| White | 38.8% | 33.3% |
| Two or more races | 4.4% | 0.0% |
| Total | 99.9% | 99.9% |
| Sex | | |
| Female | 50% | 77.8% |
| Male | 50% | 22.2% |

While recruiting student participants, I also recruited high school teacher participants. First, I obtained a list of math educators from the cooperating high school and randomly selected eight educators. The cooperating teacher made initial contact with the educators. When educators declined to participate, I randomly selected replacements and provided the names to the cooperating teacher. When educators agreed to participate, the cooperating teacher obtained a signed consent form from the participant (see Appendix A). The high school educator recruitment effort yielded eight participants.

While recruiting student and high school educator participants, I also recruited post-secondary educators. I utilized post-secondary school websites to obtain contact information for math educators in or near the city in Michigan. I emailed every potential participant and followed up with multiple emails. The post-secondary educator recruitment effort yielded eight participants from four post-secondary schools and I received their consent forms via email.

To ensure that the experimental group and the control group had similar demographics (such as gender, ethnicity). Also, if there was more than one participant representing a post-secondary school, I attempted to put those participants in different groups. I used stratified randomization to assign the 16 educator participants into a control group and an experimental group. On the day of the experimental group data collection event, one post-secondary educator could not attend due to illness. Therefore, the control group consisted of eight educators and the experimental group consisted of seven educators. Table III shows demographic information for the educator participants by group.

**Table III**

DEMOGRAPHIC SUMMARY OF EDUCATOR PARTICIPANTS

| | Control Group | Experimental Group |
|---|---|---|
| Gender | | |
| Female | 62.5% | 42.8% |
| Male | 37.5% | 57.1% |
| Race/ Ethnicity | | |
| American Indian/ Alaskan | 0% | 0% |
| Asian/ Pacific Islander | 12.5% | 14.2% |
| Black | 12.5% | 0% |
| Hispanic | 12.5% | 0% |
| White | 50.0% | 71.4% |
| Two or more races | 12.5% | 0% |
| Undeclared | 0% | 14.2% |
| Highest Degree Obtained | | |
| Bachelor | 25.0% | 0% |
| Masters | 50.0% | 71.4% |
| Doctorate | 25.0% | 28.6% |
| Prior Experience Reviewing Standardized Test Questions | | |
| No | 75.0% | 71.4% |
| Yes | 25.0% | 28.6% |
| Educators' School Level | | |
| Secondary | 50.0% | 57.1% |
| Post-secondary | 50.0% | 42.9% |
| Community college | 12.5% | 0% |
| University | 37.5% | 42.9% |
| Average Years of Teaching Experience | | |
| *M* | 13.63 | 24.14 |
| *SD* | 10.53 | 11.02 |

## F. Instruments

I describe each data collection instrument as follows.

### 1. Teacher background survey

I used the educator background survey (see Appendix B) to gather demographic data from educator participants. The purpose of the survey was to collect teacher background on education, teaching experience, and prior experience reviewing test questions. I utilized these data to evaluate the comparability of the control and the experimental groups.

2. **Test question collection**

I utilized a pre-existing collection of 20 questions designed for SAT math test (see Appendix C) from a pilot study (Trapp, 2015). These test questions are non-secure and were obtained from the College Board. I have chosen questions that are a representative sample of the mathematics content assessed on the SAT. Selected test question attributes are listed in Table IV. The second column of the table indicates the abbreviated code of alignment to the Michigan Math Standards (Michigan Department of Education, 2010). Two types of SAT questions were in the collection: 70% of the questions were 4-option multiple choice (4MC) and 30% of the questions were student produced response (SPR) where students bubble-in their numerical answer. The fourth column lists the answer key of each item. Lastly, the fifth column lists the question difficulty, based on the estimated percent of test takers who will answer the question correctly, where 20% are easy, 45% are medium, and 35% are hard. As shown in Table IV, the questions represent a variety of math content, question types, and difficulty.

In a typical question review, reviewers evaluate many questions assessing the same math concept and the questions assessing the same concept are reviewed together, test takers will encounter the easiest test questions at test beginning and the hardest test questions at the end of the test. Educators review approximately 12 questions per hour, or 100 questions per day. In this study, due to the limited time (four hours) the participants had and the additional rating processes, I suspected that the participants might not be able to review all the 20 items. Therefore I presented the questions to reviewers in a random order so that the questions reviewed had a variety of difficulty levels, math content, and test formats regardless of the number of questions reviewed.

**Table IV**

TEST QUESTION COLLECTION ATTRIBUTES

| Question number | State standard | Question type | Answer key | Question difficulty |
|---|---|---|---|---|
| 1 | HSF-TF.2 | 4MC | C | Easy |
| 2 | HSG-GPE.1 | 4MC | D | Medium |
| 3 | HSA-REI.4b | 4MC | D | Medium |
| 4 | HSF-IF.2 | 4MC | C | Hard |
| 5 | HSG-GMD.3 | SPR | 51.3 | Hard |
| 6 | HSA-CED.1 | 4MC | B | Easy |
| 7 | HSA-CED.4 | 4MC | C | Medium |
| 8 | HSA-REI.3 | SPR | $-0.75 \leq x \leq 3$ | Medium |
| 9 | HSA-REI.6 | SPR | 9/13 or .692 | Hard |
| 10 | HSF-BF.1 | 4MC | A | Medium |
| 11 | HSS-ID.7 | 4MC | C | Medium |
| 12 | HSF-LE.2 | SPR | 1/17 or .059 | Hard |
| 13 | HSA-SSE.1 | 4MC | B | Hard |
| 14 | HSA-SSE.2 | 4MC | D | Hard |
| 15 | HSA-REI.10 | 4MC | B | Medium |
| 16 | HSA-APR.3 | 4MC | A | Medium |
| 17 | HSS-ID.5 | 4MC | C | Easy |
| 18 | HSS-IC.1 | 4MC | C | Medium |
| 19 | Modeling | SPR | 14070 | Easy |
| 20 | Modeling | SPR | 73667 | Hard |

In addition to the attributes listed in Table IV, several items contain flaws based on the initial review during the pilot study. For example, question 6 does not align well to the indicated Michigan state standard. Also, questions 5 and 7 are not clear and concise and participants will likely request text revisions. I expected that all the questions would generate discussion and that some would be accepted as is, some would be accepted with revisions, and some may be rejected. The proportion of problematic questions is similar to the proportion of problematic questions in actual SAT educator question reviews.

3. **Rating form**

All the participants, including students, educators in the control group, and educators in the experimental group, each completed a rating form (see Appendix D) for each

question. Participants were given the choice to complete the form on paper, on their phone via Qualtrics, or on a provided desktop computer via Qualtrics. The first part of the rating form was completed after an individual review. Each participant indicated their personal recommendation for the test question along with comments to support their recommendation. Following a discussion about each question, the participants documented the group consensus and indicated their agreement or disagreement with the group's consensus. This rating form is the modified version of the tool I used in a pilot study (Trapp, 2015). During the pilot study, participants indicated that the former table headings were difficult to interpret during the second round of ratings. I clarified the directions and column headings on the rating form to address these concerns and have revised the format to be consistent with the procedure described later in this chapter.

4. **Educator appraisal (self-efficacy) inventory**

In addition to rating test questions, educators filled in a self-efficacy questionnaire twice during the procedure. I developed the self-efficacy questionnaire using the process described by Bandura (2006). First, I reviewed the literature regarding the judgments that educators make during question review. Based on a content analysis of the literature, I created statements for the survey. In January 2017, I piloted the survey with a group of twelve educators. Based on their responses and feedback, I revised statements that were unclear, combined statements that were too similar, and removed statements that were not applicable. The original inventory consisted of over 60 statements and the resulting questionnaire was reduced to 41 statements. In March 2017, I piloted the 41-statement survey with another group of 12 educators. Based on the responses from the second pilot, no additional revisions were required. The questionnaire is included in Appendix E.

Each statement is categorized into one of four domains that emerged during the literature review: Math Content, Construct-Irrelevance, Cognitive Psychology, and Consequences of Testing. An example question from the instrument is, "Rate your degree of confidence in evaluating a question's alignment to your state standards."

Based on the recommendations of my dissertation committee, I altered the response format of the appraisal inventory from a 0 to 100 response format to a 5-point rating scale. Participants rate each statement on a five-category format: Highly unsure, Unsure, Neutral, Confident, Highly confident. For the data analysis, I used a five-point rating scale: Highly unsure (HU) = 1, Unsure (U)= 2, Neutral (N)= 3, Confident (C)= 4 and Highly confident (HC)= 5. Bandura (2006) recommends that self-efficacy inventories do not use the term *self-efficacy*. Bandura's reasoning is the term can be mistaken for other personality aspects. Also, self-efficacy is closely connected to self-worth and including the term *self-efficacy* on the form itself may cause undesired rating effects. Therefore, I titled the form "Appraisal Inventory" and the term *self-efficacy* was not used.

5. **Student and educator feedback questionnaire**

At the end of the data collection event, each participant completed a feedback questionnaire (see Appendix F), which I used to gather feedback concerning the training provided and the process used to gather data from participants. For example, the questionnaire asks if participants felt their concerns were heard and their perception of the resulting quality of the questions reviewed. I used these data to evaluate threats to validity and to improve future data collection events.

## G. Procedure

Following the standard process used by test developers during educator question reviews, I held three separate question review events to collect data from the test taker group, the control group, and the experimental group.

### 1. Test taker question review

The nine student participants arrived, took seats, and we completed introductions. Then, I started the audio recorder and provided the question review training. Following the training, I answered any questions asked by the participants.

After answering questions, I distributed the instrument of 20 test questions from Appendix C along with the copies of the rating form from Appendix D. Beginning with the first question in the packet, I modeled the question review discussion process with the students. The modeling process consisted of the following steps: (a) individual reading of the question, (b) entering initial rating on the rating form, (c) discussing question quality and question edits which may improve question quality, (d) reaching consensus about question edits and question quality, and (e) documenting final criticism on the rating form. After modeling the process, I answered any remaining questions.

I facilitated the discussion, one question at a time. For each question, the student rating form collected the written record of student criticism while the audio recording collected the verbal discussion of student criticism. The group discussion ended when we completed the four hours allotted. The student group completed reviewing 16 questions in the four hours.

Following the conclusion of the group discussion, I collected the completed rating forms and the question packet, thanked everyone for their participation, described the final procedures,

and stopped the audio recording. Student participants completed the student-specific questionnaire and received a gift card.

I summarized the student review results in preparation for the educator review. For each question, I documented selected student quotations, a summary of the discussion, the final question recommendation, and a summary of the student's agreement with the recommendation. Lastly, if the student participants approved edits to a test question, then I provided the revised question.

2. **Educator question review**

I used an experimental research design to examine the impact of test taker criticism on educators' question review self-efficacy. I recruited 16 educators. After stratifying the educators in sub-groups according to gender, race/ethnicity, and the school level that they were teaching, I randomly assigned half of them to an experimental group and the other half to a control group. One educator from the experimental group was unable to participate due to illness and therefore the experimental group was composed of seven educators. Table V displays the major procedures of the experiments. As shown in Table V, the control group and the experimental group differed in their treatment: the experimental group had students' criticism of the test items available for their group discussion while the control group did not have that information. Details of the procedures are as follows.

**Table V**

MAJOR PROCEDURES OF THE TWO GROUPS

| Group | Pretest | Treatment | Posttest |
|---|---|---|---|
| Control | Self-efficacy survey | Group discussion based on educator comments only | Self-efficacy survey<br>Test review results |
| Experimental | Self-efficacy survey | Group discussion based on both educator comments and test taker criticism | Self-efficacy survey<br>Test review results |

a. **Control group**

Eight educator participants arrived, were seated, and we completed introductions. Then, I started the audio recorder and began the training. Following the training, I provided the participants an opportunity to ask questions and answer their questions. Then, I distributed the instrument of 20 test questions from Appendix C and the rating form from Appendix D to the participants. Using the first question in the instrument, I modeled the question review process. The modeling process consisted of the following steps: (a) individual reading of the question, (b) entering initial rating on the rating form, (c) discussing question quality and question edits which may improve question quality, (d) reaching consensus about question edits and question quality, and (e) documenting final criticism on the rating form. After modeling the process, I provided participants another opportunity to ask questions and answered those questions. Then, I facilitated the participants through the remaining questions, one item at a time. The control group finished reviewing the first 10 questions in the packet in the four hours allotted. After reviewing the tests, the participants completed the self-efficacy inventory posttest and the end survey on Qualtrics. Before completing the appraisal inventory, I reminded the educators to consider each statement individually and thoughtfully, and that there is no correct answer. Finally participants were given the feedback questionnaire. After participants completed the surveys on Qualtrics, I provided each participant with a gift card.

b. **Experimental group**

The procedures for the experimental group were identical to the procedures for the control group with the exception of the treatment. After the educators completed individual review of a test question and their individual ratings, I provided them with a written summary of test taker criticism which consisted of: (a) initial and final student ratings, (b) students' written and

verbal criticism transcribed, and (c) the test question with the student approved edits implemented. The written summary of test taker criticism is provided in Appendix G. Then, I facilitated the group discussion using the same process as the control group and the educators documented their final rating and final criticism of the question on the rating form. The experimental group only finished reviewing the first five questions in the packet in the four hours allotted. Similar to the control group, after completing the four-hour test review, the experimental group completed the self-efficacy inventory, feedback questionnaire, and received a gift cards.

Table VI summarizes the procedures completed by the two groups and the corresponding instruments used when applicable. Notice that the procedures taken by the two groups were almost identical except that the experimental group received a written summary of test taker criticism between their individual rating and group discussion.

**Table VI**

EXPERIMENTAL PROCEDURES AND INSTRUMENTS

| Procedures | Instrument |
|---|---|
| Before question review | |
| • Completing demographic survey | Educator profile (Appendix B) |
| • Completing self-efficacy appraisal inventory | Educator appraisal (self-efficacy) inventory (Appendix E) |
| During question review | |
| • Receiving question review training | |
| • Reviewing a question individually | Test question collection (Appendix C) |
| • Entering initial rating | Rating form (Appendix D) |
| • Receiving and reviewing a written summary of test taker criticism (Experimental group only) | Summary of Test Taker Criticism (Appendix G) |
| • Discussing question quality and question edits | |
| • Reaching consensus about question status and/or question edits | |
| • Entering final comments on the rating form | Rating form (Appendix D) |
| After question review | |
| • Completing self-efficacy appraisal inventory | Educator appraisal (self-efficacy) inventory (Appendix E) |
| • Completing feedback questionnaire | Educator feedback questionnaire (Appendix F) |

# IV. RESULTS

I conducted several analyses to answer the two research questions stated in Chapter I. First, I compared two groups' educator self-efficacy using split-plot analysis of variance (ANOVA) and analysis of covariance (ANCOVA). Next, I used mixed methods to compare two groups' educators' test question review quality. Prior to conducting these analyses, I examined whether sufficient reliability and validity evidence existed to support the use of the appraisal inventory in these analyses.

A. **Analyses Addressing Research Question 1**: To what degree does educator exposure to test taker criticism of questions result in an increase in educator's self-efficacy regarding their question review judgments?

To answer research question 1, I applied the ANCOVA and the split-plot ANOVA procedures to the pretest and posttest appraisal inventory ratings. The ANCOVA procedure investigated whether the appraisal inventory posttest scores were significantly different across groups while controlling for pretest scores. The split-plot ANOVA procedure investigated whether educator self-efficacy significantly increased from pretest to posttest for each group of educators. Prior to conducting ANCOVA and split-plot ANOVA on the data, I investigated the reliability of the appraisal inventory scale and I examined the assumptions for ANCOVA and split-plot ANOVA.

1. **Reliability of the appraisal inventory scale**

Eight participants in the control group and seven participants in the experimental group each responded to the appraisal inventory as both a pretest and a posttest. I analyzed the reliability of the administration data by domain and group and the results are provided in Table VII. Nunnally (1978) suggests that alpha coefficients of 0.90 and above are tolerable for high stakes testing and values of 0.80 and above are appropriate when investigating correlations

between different variables. The reliability analysis revealed that all the alpha coefficients are 0.80 or higher, except for one, control group's responses for the consequences of testing domain. Given that the appraisal inventory is not a high stakes test, these results suggest that the administration data are reliable, but caution should be exercised when carrying out analyses with the control group's responses for the consequences of testing domain. The alpha coefficient for the overall inventory scores is high, all above .90 for the control group, experimental group, and all participants.

**Table VII**

APPRAISAL INVENTORY RELIABILITY RESULTS

| Domain | Sample statement | Number of statements | Group | Cronbach Alpha |
|---|---|---|---|---|
| Math content | Rate your degree of confidence in evaluating a question's content based on the question's alignment to my state standards. | 11 | Control<br>Experimental<br>Combined | 0.87<br>0.88<br>0.87 |
| Construct-irrelevance | Rate your degree of confidence in evaluating a question's unfairness to test takers as a result of unfairness due to a context impacting test taker motivation. | 16 | Control<br>Experimental<br>Combined | 0.84<br>0.97<br>0.94 |
| Cognitive psychology | Rate your degree of confidence in evaluating the appropriateness of a question that differentiates between test takers who use weak problem-solving methods (e.g. trial and error) and test takers who use strong methods (e.g. standard algorithms, mathematical reasoning) based on the mental processes that test takers are likely to utilize. | 9 | Control<br>Experimental<br>Combined | 0.88<br>0.95<br>0.92 |
| Consequences of testing | Rate your degree of confidence in evaluating a test question's appropriateness based on the intended or unintended consequence that the question may change test preparation efforts of test takers | 5 | Control<br>Experimental<br>Combined | 0.72<br>0.90<br>0.84 |
| Entire appraisal inventory | | 41 | Control<br>Experimental<br>Combined | 0.94<br>0.98<br>0.97 |

Besides the alpha coefficient, I also calculated the Cronbach's alpha value if a statement was deleted and the corrected item-total correlation for each statement. The reliability of the inventory did not greatly increase or decrease with the removal of any single inventory statement. Corrected item-total correlations for all statements were .40 or higher with the exception of the second statement in the math content domain:

> Math content, statement 2: Rate your degree of confidence in evaluating a question's content based on the question's mathematical precision.

This statement is important to content validity, so I decided to retain the statement even though the corrected item-total correlation was low.

After determining that the inventory data collected are reliable, I generated a composite score for each participant by averaging their ratings for the 41 statements. Therefore, the inventory scale has a minimum value of 1 and a maximum value of 5. I calculated pretest and posttest self-efficacy composite scores by averaging the ratings for all 41 statements on each occasion. I also calculated domain scores by averaging the ratings for the corresponding statements measuring each domain. Table VIII shows these results further summarized by appraisal inventory domain and test administration (pretest and posttest). The mean self-efficacy scores increased from pretest to posttest across all the domains.

**Table VIII**

APPRAISAL INVENTORY DOMAIN SCORE DESCRIPTIVE STATISTICS

| | Number of | Appraisal Inventory Administration | | | | | |
|---|---|---|---|---|---|---|---|
| | | Pretest | | Posttest | | Total | |
| Domain | questions | *M* | *SD* | *M* | *SD* | *M* | *SD* |
| Math content | 11 | 3.32 | 1.07 | 3.96 | 0.81 | 3.64 | 1.00 |
| Construct-irrelevance | 16 | 3.70 | 0.72 | 4.10 | 0.67 | 3.90 | 0.72 |
| Cognitive psychology | 9 | 3.53 | 0.79 | 4.04 | 0.76 | 3.79 | 0.81 |
| Consequences of testing | 5 | 3.45 | 0.74 | 4.04 | 0.74 | 3.75 | 0.80 |
| Total | 41 | 3.53 | 0.86 | 4.04 | 0.74 | 3.79 | 0.84 |

The appraisal inventory domains ask participants to self-report their confidence in reviewing different question aspects, which may be interrelated. To investigate the relationship, I calculated correlation coefficients between domains. Table IX shows that all correlations are statistically significant within each inventory administration. These results suggest that the domain scores are all positively related.

**Table IX**

CORRELATION COEFFICIENTS AMONG APPRAISAL INVENTORY DOMAINS (*N*= 15)

| | Administration & Domain | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 |
|---|---|---|---|---|---|---|---|---|---|
| 1 | Pre math content | - | | | | | | | |
| 2 | Pre construct-irrelevance | .77* | - | | | | | | |
| 3 | Pre cognitive psychology | .84* | .64* | - | | | | | |
| 4 | Pre conseq. of testing | .89* | .84* | .82* | - | | | | |
| 5 | Post math content | .47 | .56* | .44 | .32 | - | | | |
| 6 | Post construct-irrelevance | .56* | .63* | .57* | .60* | .58* | - | | |
| 7 | Post cognitive psychology | .50 | .54* | .56* | .45 | .68* | .86* | - | |
| 8 | Post conseq.of testing | .46 | .48 | .31 | .45 | .52* | .79* | .76* | - |

* $p < .05$

2. **Examination of ANCOVA and split-plot ANOVA assumptions**

a. **ANCOVA.** ANCOVA requires the following assumptions: (a) the dependent and covariate variables are measured on a continuous scale (b) observations are independent (c) no significant outliers (d) the independent variable residuals are approximately normally distributed (e) homogeneity of variances (f) homoscedasticity (g) the relationship between the dependent variable and the covariate is linear and (h) homogeneity of regression slopes (Huitema, 2011). The results below describe the analyses I used to evaluate these assumptions.

The dependent variable, posttest, and the covariate variable, pretest, were measured on a scale with a minimum value of 1 and a maximum value of 5. A respondent's scale score was

obtained by averaging their responses to the 41 statements. Using the composite score constitutes a continuous scale (Harpe, 2015).

The research design ensured that all ratings were independent. I monitored the participants while they completed the posttest to ensure there was no interaction. All participants completed the pretest at least one week in advance of the posttest and participants did not interact with each other during the pretest or posttest. Also, the members of the control group and the experimental group were mutually exclusive and they participated the study separately.

There were no significant outlier scores in either the pretest or the posttest data. All pretest scores were within two standard deviations of the pretest score mean. All posttest scores were within two standard deviations of the posttest score mean.

The Kolmogorov-Smimov test, the Shapiro-Wilk test and the Normal Q-Q Plots all showed that the normality was met. I used Levene's test of equality of error variances to test homogeneity of variance and found that the assumption was met. To evaluate homoscedasticity, I generated a scatterplot of the standardized residuals versus the dependent variable. The variance of the residual terms appears to be constant, therefore this assumption was met.

To determine whether the relationship between the dependent variable and the covariate is linear, I generated a scatterplot of the covariate variable versus the dependent variable, which indicated a linear relationship.

I investigated the assumption of homogeneity of regression slopes first by generating a scatterplot of the covariate variable versus the dependent variable by group and then adding regression lines by group. The scatterplots show that the slopes of the regression lines do not intersect within the domain of the graph, but the slopes are not parallel. Using post self-efficacy score as the dependent variable, I fitted an ANOVA model of the independent variable, group, and

the covariate, pretest, as well as all an interaction between group and pretest. The interaction between the pretest score and group was not significant and therefore the homogeneity of regression slope assumption was met.

b. **Split-plot ANOVA**. Split-plot ANOVA requires two additional assumptions besides normality, independent observations, and homogeneity of variance examined above: (a) sphericity, which means that the variances are equal to each other and covariances are equal to each other within each cell of the between-subjects factor and (b) equal variance-covariance matrices, that is, the variance and covariance matrices are the same across the cells of the between-subject factor. The assumption of sphericity is required when there are three or more repeated measures. This study included two repeated measures, pretest and posttest, so sphericity does not apply. I used Box's M test of equality of covariance matrices and found that the assumption was met.

3. **ANCOVA results and effect size**

I conducted a one-way ANCOVA analysis to answer research question 1. The independent variable was group and the covariate was the pretest total score on the appraisal inventory. The dependent variable was posttest total score and the control group included eight participants and the experimental group included seven participants. I performed the analysis with the SPSS statistical package. Table X presents the means and standard deviation of the pretest and posttest self-efficacy scores of the two groups. ANCOVA shows a significant effect of group (control or experimental) on posttest total score after controlling for pretest total score, $F(1, 12) = 11.42$, $p = .005$, *partial* $\eta^2 = .49$. Overall, these data suggest that exposure to test taker criticism during a test question review results in educators reporting a higher question review self-efficacy

than if they had not been exposed to test taker criticism after controlling for their pretest self-efficacy.

**Table X**

APPRAISAL INVENTORY MEANS BY GROUP

| Group | Pretest *M* (*SD*) | Posttest *M* (*SD*) | Adjusted Posttest *M* (*SD*) |
|---|---|---|---|
| Control | 3.54 (.40) | 3.88 (.40) | 3.87 (.37) |
| Experimental | 3.52 (.50) | 4.22 (.59) | 4.23 (.37) |

4. **Split-Plot ANOVA results and effect size**

I conducted a split-plot ANOVA to compare the differences in the pre- and posttest total self-efficacy ratings by groups. There was a significant main effect of test administration on self-efficacy ratings $F(1,13) = 29.56$, $p < .001$, *partial* $\eta^2 = .70$. There was no significant effect for group, suggesting that the self-efficacy ratings for the control group and the experimental group were generally the same when both pretest and posttest are considered. Also, there was no significant interaction between administration and group. Figure 1 shows the change in mean educator self-efficacy ratings. The two groups started with similar pretest self-efficacy but the experimental group scored higher than the control group on posttest self-efficacy, which is consistent with the ANCOVA results.

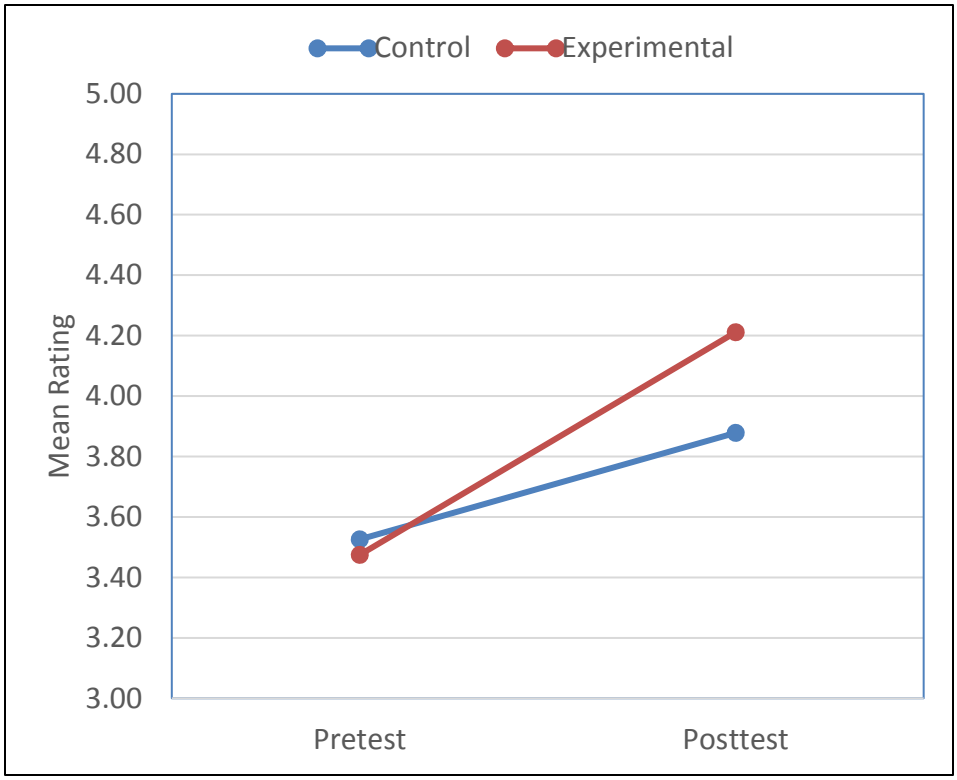



Figure 1. Split-plot ANOVA: total score.

5. **Preliminary analyses by appraisal inventory domain**

As the four self-efficacy domains focus on different aspects of question reviews, I examined the impact of test taker's criticism on educators' appraisal inventory scores by domain as well. I generated domain scores for each participant by averaging their ratings for the associated domain statements. I compared the differences in the pre- and posttest self-efficacy ratings for each appraisal inventory domain using split-plot ANOVA analyses with a Bonferroni adjustment. Prior to conducting the analyses, I tested the assumptions needed for each domain score. Most assumptions were met. However, the construct-irrelevance posttest scores and the consequences of testing pretest scores both failed the test of normality. Therefore, I will consider these results preliminary and I will not draw any decisive conclusions. Future research may replicate this study and test these results with a larger sample size or with non-parametric tests.Table XI shows the

results of the split-plot ANOVA analyses and Figures 2 through 5 present the domain score mean plots from pretest to posttest for the two groups. There was a significant main effect of administration for all domains which means that participants significantly increased all domain self-efficacy scores from pretest to posttest. There was no significant effect for group in any of the domains, suggesting that self-efficacy ratings for the control group and the experimental group were generally the same when pretest and posttest are considered together. The interaction between test administration and group was significant for the construct-irrelevance domain and the consequences of testing domain. This result indicates that the two groups' rating change from pretest to posttest ratings did not differ significantly on math content and cognitive psychology, but the rating change differed significantly on construct-irrelevance and consequence of testing.

**Table XI**

F TEST RESULTS AND SIGNIFICANCE BY DOMAIN

| | Effect tested | | |
|---|---|---|---|
| Domain | Administration | Group | Administration * Group |
| Math content | $F(1,13) = 19.16$, $p = .001$, *partial* $\eta^2 = .60$ | $F(1,13) = .63$, $p = .441$, *partial* $\eta^2 = .05$ | $F(1,13) = .01$, $p = .935$, *partial* $\eta^2 = .60$ |
| Construct-irrelevance | $F(1,13) = 23.73$, $p < .001$, *partial* $\eta^2 = .65$ | $F(1,13) = .74$, $p = .405$, *partial* $\eta^2 = .05$ | $F(1,13) = 9.87$, $p = .008$, *partial* $\eta^2 = .43$ |
| Cognitive psychology | $F(1,13) = 13.00$, $p = .003$, *partial* $\eta^2 = .50$ | $F(1,13) = 1.15$, $p = .303$, *partial* $\eta^2 = .08$ | $F(1,13) = .15$, $p = .708$, *partial* $\eta^2 = .01$ |
| Consequences of testing | $F(1,13) = 24.55$, $p < .001$, *partial* $\eta^2 = .65$ | $F(1,13) = .09$, $p = .771$ *partial* $\eta^2 = .015$ | $F(1,13) = 7.38$, $p = .018$, *partial* $\eta^2 = .36$ |

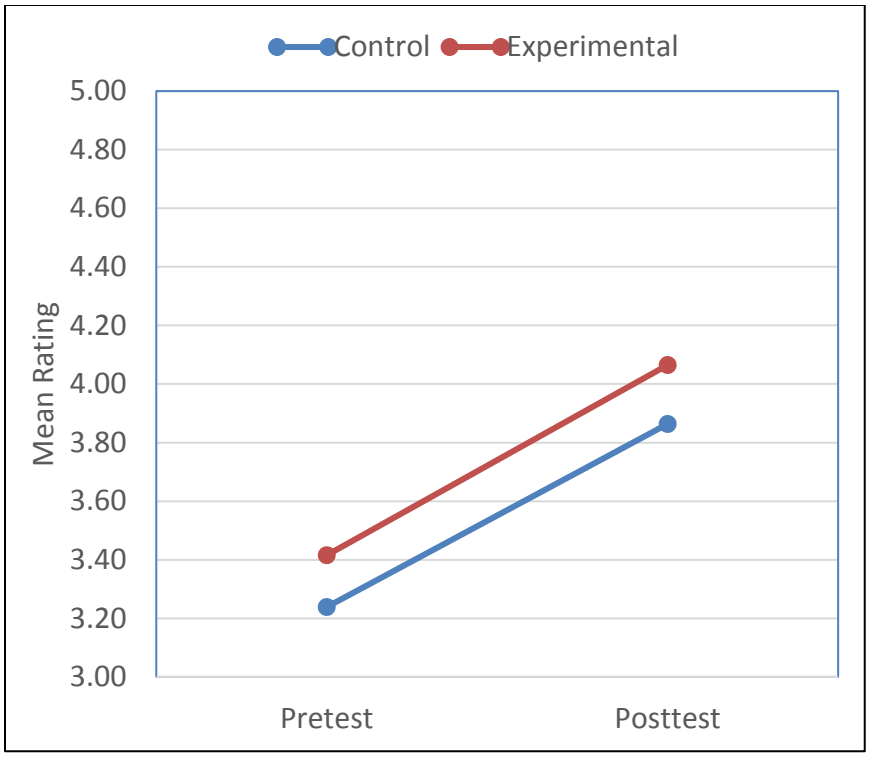


Figure 2. Split-plot ANOVA: math content domain score.

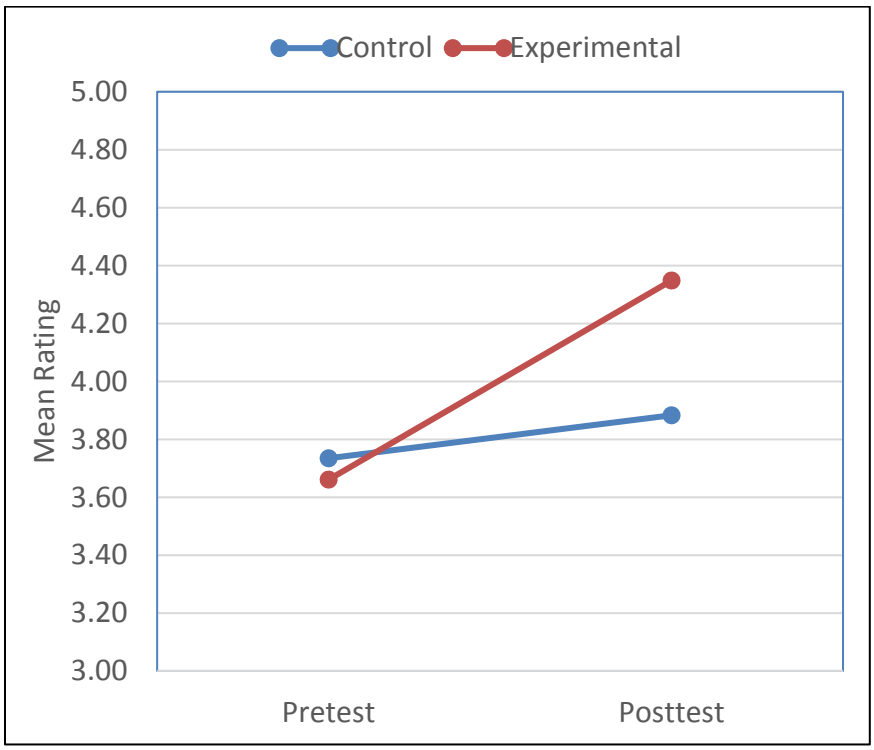


Figure 3. Split-plot ANOVA: construct-irrelevance domain score.

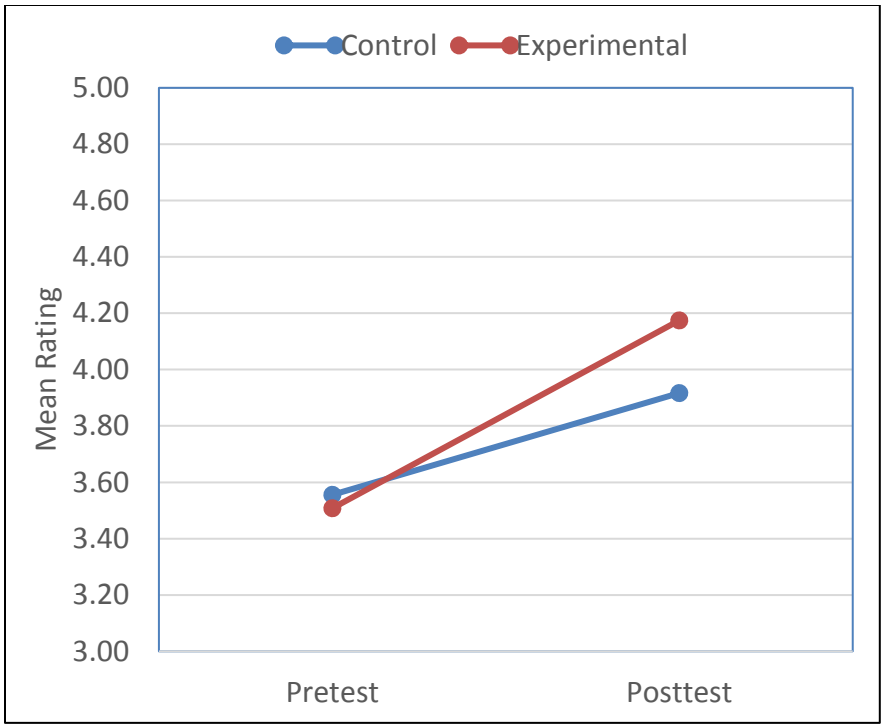


Figure 4. Split-plot ANOVA: cognitive psychology domain score.

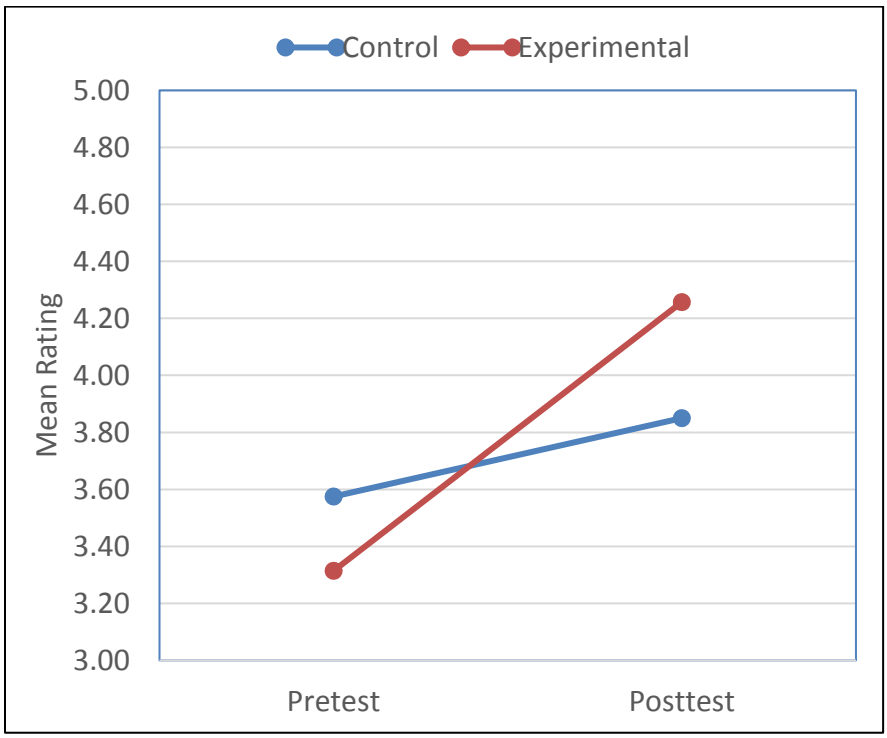


Figure 5. Split-plot ANOVA: consequences of testing domain score.

I also conducted preliminary ANCOVA analyses by domain. The independent variable was group and the covariate was the pretest total score for each appraisal inventory domain. The dependent variable was posttest domain score and the control group included eight participants and the experimental group included seven participants. Table XII presents the means and standard deviations of the pretest and posttest self-efficacy scores of the two groups by domain, and the ANCOVA results for group effect. Except for the math content domain, the experimental group significantly outperformed the control group on almost on almost all domains after controlling for the corresponding pretest scores.

These results suggest that most of the domain-level data behaved similarly to the total self-efficacy data and the experimental group pretest scores were significantly higher than the control group, after controlling for pretest scores.

**Table XII**

ANCOVA RESULTS BY DOMAIN

| Domain | Group | Pretest *M* (*SD*) | Posttest *M* (*SD*) | Adjusted Posttest *M* (*SD*) | F test | *partial* $\eta^2$ |
|---|---|---|---|---|---|---|
| Math content | Control | 3.24 (.61) | 3.86 (.46) | 3.92 (.51) | $F(1, 12) = 3.13$ | .21 |
| | Experimental | 3.42 (.42) | 4.06 (.63) | 4.02 (.51) | $p = .102$ | |
| Construct-irrelevance | Control | 3.73 (.38) | 3.88 (.33) | 3.86 (.33) | $F(1, 12) = 15.73$ | .57 |
| | Experimental | 3.66 (.52) | 4.34 (.63) | 4.38 (.34) | $p = .002$ | |
| Cognitive psychology | Control | 3.56 (.46) | 3.91 (.54) | 3.90 (.54) | $F(1, 12) = 6.10$ | .34 |
| | Experimental | 3.50 (.66) | 4.17 (.73) | 4.19 (.54) | $p = .029$ | |
| Conseq. of testing | Control | 3.58 (.43) | 3.85 (.54) | 3.77 (.48) | $F(1, 12) = 6.93$ | .37 |
| | Experimental | 3.31 (.56) | 4.26 (.59) | 4.35 (.48) | $p = .022$ | |

B. **Analyses Addressing Research Question 2**: To what degree does educator exposure to test taker criticism of questions result in a higher quality question review during the question review process?

Test taker criticism was provided to the experimental group during the question review process. After the experimental group participants rated each question individually, I provided a written summary of test taker criticism for that question to each participant. The summary included the recommendation from the test takers (accept as is, accept with revisions, or reject) along with any recommended question edits. Also, selected quotes from students were provided, a summary of the discussion, and a summary of the agreement that test takers had with the recommendation. The test taker criticism is provided in Appendix G.

I employed a mixed method data analysis to answer question 2 because both the quantitative data and qualitative data were required to illuminate the relationship between test taker criticism exposure and test review quality. The quantitative analysis indicated the participant's finality of their ratings on the appraisal inventory and the rating form while the qualitative data showed the participant's hesitation, hedging, and rationales. For example, the transcripts reveal that a participant confidently state a concern about a particular question aspect, followed by uncertainty whether a change is required to address the concern. As the quantitative and qualitative data illuminate the same variable, test review quality, with different emphases, I collected and analyzed both types of data concurrently. Creswell (2002) refers to this methodology as triangulation. I utilized qualitative data from the transcribed question review event and responses to the open-ended questions on the rating form. Quantitative data included ratings from the appraisal inventory, initial question ratings, final question ratings, as well as descriptive statistics resulting from counting the frequency of the codes appearing in the transcripts.

In the following paragraphs, I briefly summarize the data collection event results, the level of agreement with the recommendations, the coding process and interrater reliability results, and a summary of the codes identified. Then, I summarize the results of the triangulation data analysis by major themes that emerged. Under each theme, I mix codes, quotes, anecdotes, and descriptive statistics together to describe the differences in test review quality of the control and experimental groups. In the next chapter, I utilize the results of these analyses to answer research question 2.

The experimental research design of this study partially relies on group experiences being as similar as possible. Both groups participated in the data collection event for four hours, but the control group reviewed the first 10 questions while the experimental group only reviewed the first five questions. These differences in the number of questions reviewed did not seem to impact the quantity of qualitative data obtained. Details are provided below.

1. **Question review results**

A test question review is typically summarized by the final recommended question status: accepted as is, accepted with recommended revisions, or reject. Each of these ratings is distinct and often reveal strengths and weaknesses of participants. For example, a participant who enters an initial rating of *Accept As Is* may have missed one or more flaws in the question. Participants began the question review process by reviewing each question individually and reporting their recommended status. The interrater reliability of each group's recommended status was analyzed with interclass coefficients at 95% confidence intervals for *k* raters based on a single-rating, absolute-agreement, 2-way mixed-effects model (Koo & Li, 2016). Table XIII shows the results of the interrater reliability analysis. The interclass coefficients reveal poor agreement within each group before group discussion. Due to the small sample size, the 95% confidence intervals are large and are indicative of poor reliability (less than .50) and the intervals for the control and

experimental groups contain the lower end of what is considered moderate agreement (values between 0.5 and 0.75) (Koo & Li, 2016). Educator participants in question reviews typically have low interrater reliability of initial recommendations, which is expected because different educators have different strengths and identify different issues based on the personal experience and biases. The low interrater reliability in this study may also reflect the fact that most participants did not have prior question review experience.

**Table XIII**

INITIAL RECOMMENDATIONS INTERCLASS COEFFICIENTS BY GROUP

| Group | *k* | Interclass coefficient | 95% confidence interval | |
|---|---|---|---|---|
| | | | Lower bound | Upper bound |
| Test takers | 9 | 0.17 | 0.06 | 0.39 |
| Control | 8 | 0.28 | 0.09 | 0.62 |
| Experimental | 7 | 0.06 | –0.07 | 0.59 |

After each group reached consensus on a question, I documented the final recommendation. Table XIV summarizes the recommendations by group and the resulting questions from each group are provided in Appendix I.

**Table XIV**

FINAL RECOMMENDATIONS BY GROUP

| Question | Test takers | Control | Experimental |
|---|---|---|---|
| 1 | Accept with revisions | Reject | Accept with revisions |
| 2 | Accept with revisions | Accept as is | Accept with revisions |
| 3 | Accept as is | Accept with revisions | Accept as is |
| 4 | Accept with revisions | Accept with revisions | Accept with revisions |
| 5 | Accept with revisions | Accept with revisions | Accept with revisions |
| 6 | Accept as is | Reject | |
| 7 | Accept with revisions | Accept with revisions | |
| 8 | Accept with revisions | Accept with revisions | |
| 9 | Accept as is | Accept with revisions | |
| 10 | Accept with revisions | Accept with revisions | |
| 11 | Accept as is | | |
| 12 | Accept with revisions | | |
| 13 | Reject | | |
| 14 | Reject | | |
| 15 | Accept as is | | |
| 16 | Accept as is | | |

After each group reached consensus, they completed their rating form and indicated whether they agreed with the final consensus reached. Each control group participant agreed with the final consensus on all questions except for question 7. On the final section of their rating form, Fifty percent of the control group suggested that the group's consensus for number 7 was too hasty and the dissenters thought that the question could have been accepted with some additional revisions if we had spent more time on it. In the experimental group, all participants agreed with the final consensus for questions 1, 4, and 5. There was one dissenter each on questions 2 and 3, and both dissenters disagreed that the question was a good match to the Michigan state standard.

As shown in Table XIV, each group recommended to resolve issues in different ways. There is no correct way to resolve an issue and a high-quality recommendation prevents a flawed question from being accepted. The control group and the experimental group reached the same recommended status for two of the five common questions. The control group and the student

group reached the same recommended status for five out of the ten common questions. In contrast, the experimental group reached the same recommended status as the test taker group for all five questions. This summary suggests that the experimental group was influenced by the test taker criticism in some way and tended to agree with students' criticism.

Questions 1 to 5 were reviewed by all the three groups. In the following paragraphs, I discuss the issues uncovered in questions 1 through 5 and how the three groups addressed the issues.

a. **Question 1**: If $\cos \theta = 0.5$, which of the following is also equal to 0.5?

A. $\sin \theta$
B. $\tan \theta$
C. $\sin (90 - \theta)$
D. $\tan (90 - \theta)$

The first question had two issues related to math content. First, the question was not aligned to the state standard because test takers who do not understand the unit circle could answer the question correctly based on other knowledge and the state standard implies that this skill should be assessed in the context of angle measures in radians, not degrees. Second, answer choices C and D contain the value '90', which is intended to be in degrees, '$90^{o}$', but is not labeled accordingly. The test takers did not recognize these issues and did not propose changes to the question. The control group identified both issues and recommended to reject the question rather than resolve the issues with question edits. The experimental group also identified both issues and they revised the question. The resulting questions did not contain any issues. No experimental group participant mentioned the test taker comments for question 1, likely because the test takers did not detect any issues with the question. Both the control group and the experimental group resolved the issues appropriately, although the two groups had different recommendations. The control group made a quick decision to reject the question which saved

them time to review more questions. The experimental group took time to revise the question which resulted in a viable, approved question.

b. **Question 2:** In the *xy*-plane, what is the radius of the circle with equation $x^2 - 6x + y^2 + 4y = 8$?

A. $\sqrt{3}$
B. $2\sqrt{2}$
C. $\sqrt{13}$
D. $\sqrt{21}$

The second question did not have any math content issues, but the question used unfamiliar wording that had implications for construct irrelevance and consequences of testing. The use of the phrase, "In the *xy*-plane" appears on some SAT test questions to aid in question defensibility. Test takers with advanced mathematical knowledge may challenge the mathematical accuracy of questions, therefore specifying the type of plane eliminates ambiguity. However, the use of "In the *xy*-plane" may introduce construct irrelevance for some test takers. The test takers did not discuss this issue related to construct irrelevance, but they did discuss a consequences of testing issue connected to their opportunity to learn. Specifically, test takers were concerned about the time lag since they learned the skill and the minimal amount of class time that was spent learning the skill. The test takers recommended to insert a standard equation in the question that would remind them of what skill was required without telling them how to implement the skill. This recommendation resolved the issue related to construct irrelevance. The control group identified the issues related to construct irrelevance and consequences of testing, but recommended to accept the question without any changes. The experimental group also identified the issues related to construct irrelevance and consequences of testing. One experimental group participant referred to the test taker criticism for question 2 regarding the lag time since they learned this skill, and the experimental group recommended to implement the test taker's recommendations which resolved all question issues.

c. **Question 3**: If the equation $x^2 - 25x + c = 0$ has only one real solution, what is the value of c?

A. 5

B. $\frac{25}{2}$

C. $\frac{125}{4}$

D. $\frac{625}{4}$

The third question did not have any issues. The test takers accepted the question as is. The control group discussed the interaction of the alignment and the calculator portion of the test, and if the fractions may cause construct irrelevance in the form of anxiety. They ultimately recommended to move the question to the no calculator portion and change to the question content to reduce the difficulty. The experimental group also discussed the interaction of the alignment and the calculator portion of the test, and how the fractional answer choices were increasing the question difficulty. One experimental group participant briefly stated that the test taker criticism indicated they were OK with the question. After discussing several possible revisions, the experimental group felt that the original question was appropriate without changes, which was consistent with the test taker's criticism, and recommended no changes.

d. **Question 4**: If $f(x) = \frac{x(x-1)}{2}$, which of the following is true about $f(x-1)$?

A. $f(x - 1) = 1 - x + f(x)$

B. $f(x - 1) = f(x) - f(1)$

C. $f(x - 1) = (x - 1)f(x)$

D. $f(x - 1) = \frac{f(x) - f(1)}{2}$

The fourth question had two math content issues. First, the alignment to the state standard was not tight because the format of the answer choices requires students to complete additional mathematical steps that are not related to the concept being assessed. Also, the answer key is *A* and it was mislabeled as *C*. The test takers discussed the question difficulty without determining that the difficulty stemmed from the weak alignment. Also, the test takers did not

identify that the incorrect key was marked. The test taker group recommended revisions to the question that unintentionally resolved both issues. Both the control group and the experimental group discussed the uncertain alignment and the incorrect key. The control group revised the question to address both issues, and the result was very similar to the test taker's recommendation. The test taker's criticism for question 4 was mentioned by the experimental group multiple times, including praise from educator 6 who said, "Yes, I'm impressed with the student suggestions." The experimental group recommended revisions to some of the answer choices, but not all as the test takers had recommended.

e. **Question 5**: A gas tank manufacturer makes its tanks in the shape of a cylinder with hemispheres on both ends as shown.

The diameter of the cylinder and hemisphere is always 40 inches, but the length of the cylinder can vary. If the manufacturer wants to make a tank that has a volume of 100,000 cubic inches, how long should the cylinder part of the tank measure, to the closest tenth of an inch?

The fifth question had issues related to alignment, clarity, incorrect key, construct irrelevance, and consequences of testing. The alignment issue is a result of the phrase, "but the length of the cylinder can vary" in the second sentence. Some participants felt that this phrasing was distracting test takers from the concept assessed. All groups indicated the question was confusing because of the order in which the information was presented and the poor sentence structure. Also, everyone thought that the graphic was misleading and if it wasn't fixed, then certain subgroups of students would have an advantage on this question over other students.

Lastly, when solving this question students need to use the constant π. Students may approximate this value with 3.14, or they may use the value on their calculator key if they brought a calculator. The educators were concerned whether this test question was requiring students to use the calculator key or the approximation because that would likely impact test preparation and may be contrary to what teachers are currently doing in their classrooms. Concerns about test preparation and classroom teaching impacts are related to consequences of testing. The test takers did not identify the incorrect key and focused on resolving clarity issues. All remaining issues stated above remained in the test taker's proposed revision. The control group discussed all issues and recommended several revisions, including converting the question from SPR format to 4MC format. However, their final recommended question did not resolve the alignment issue. The experimental group also discussed all issues and recommended several revisions. The experimental group commended on the test taker criticism once: one participant was surprised that no test takers mentioned the common error of entering the diameter instead of the radius into the volume equations. The final experimental group recommendation did not resolve the consequences of testing issue.

As shown in table XV, both the control group and the experimental group identified the question issues while the test takers missed the majority of question issues.

**Table XV**

QUESTION ISSUE IDENTIFICATION

| Question | Issue Type | Test Taker Group | Control Group | Experimental Group |
|---|---|---|---|---|
| 1 | Alignment | | ✓ | ✓ |
| | Content error | | ✓ | ✓ |
| 2 | Construct-irrelevance | | ✓ | ✓ |
| | Consequences of testing | ✓ | ✓ | ✓ |
| 3 | None | N/A | N/A | N/A |
| 4 | Alignment | | ✓ | ✓ |
| | Key | | ✓ | ✓ |
| 5 | Alignment | | ✓ | ✓ |
| | Clarity | ✓ | ✓ | ✓ |
| | Key | | ✓ | ✓ |
| | Construct-irrelevance | | ✓ | ✓ |
| | Consequences of testing | | ✓ | ✓ |

Besides questions 1 to 5, the test taker group also reviewed questions 6 through 16 and the control group reviewed questions 6 through 10. During the discussions for question 6, the control group discussed construct-irrelevance and clarity issues and recommended to reject the question while the test takers discussed the question alignment and accepted the question as is. When discussing question 7, the test takers missed a content error in the acceleration equation but the control group identified it and addressed it. These discussions and recommendations are similar to the question 4 discussion and recommendations for both groups. Both the test taker and the control group proposed edits to clarify question 7 and recommended to accept the question with their revisions. The separate discussions and resulting recommendations for questions 8 and 10 had many parallelisms to the discussions for questions 4 and 5, respectively. There were no content errors in either question, and the themes discussed and resulting recommendations were consistent with the previously reviewed questions. Lastly, question 9 did not have any content issues but the

phrasing of the question was awkward. The control group recommended to revise the question and the test takers accepted the question as is.

In their discussions of questions 6 through 10, the test takers missed content errors but helped identify problematic wording and difficulty that may have stemmed from construct-irrelevance issues. The control group continued to identify and resolve issues related to the four domains of the appraisal inventory in their review of questions 6 through 10. Based on experimental group's review of questions 1 to 5, I suspect that if the experimental group had reviewed questions 6 through 10, they would likely have identified the same issues as the control group and resolved them with a slightly different, but still high-quality, recommendation.

2. **Coding process and resulting themes**

After transcribing the control group and experimental group discussions, I reviewed the transcripts and corrected any errors in preparation for coding. Then, I recruited a math test development expert to assist me with coding. This researcher has four years of high school math teaching experience, four years of experience writing standardized math test questions, two years of math test development experience and has previously assisted me with collecting and analyzing data for research studies conducted by the College Board. The researcher is trained in coding and was a reliable coder on a previous project.

The researcher and I separately performed a detailed, line-by-line open coding (Strauss & Corbin, 1998) on the transcription of the control group's discussion of the first two questions. During open coding, we remained flexible in the way that we categorized and described phenomena found in the text. We both completed our first review without structure to see what themes emerged. After we met and discussed our themes, we decided to use the question review framework, including four major domains: math content, construct-irrelevance, cognitive

psychology, and consequences of testing. Then, we compared our resulting themes, reconciled any differences, and updated the theme names. Using the revised list of theme names, we separately coded the control group's discussion of the first four questions.

I performed interrater reliability analysis to determine the degree that coders consistently assigned the themes to statements in the transcripts. The two coders aligned each participant statement to one theme, multiple themes, or no themes. The marginal distributions of the theme assignments did not indicate prevalence or bias problems, suggesting that kappa (Cohen, 1960) was an appropriate index (Eugenio & Glass, 2004). After coding the first four questions, the percent agreement was 0.86 and the Cohen's kappa was 0.62. After comparing discrepancies and discussing differences, we recoded the four questions and the percent agreement was 0.89 and Cohen's Kappa was 0.69. These results indicate that the raters had substantial agreement when assigning statements to the themes (Landis & Koch, 1977). Due to time and resources available, I dismissed the other researcher and coded the remaining statements.

I utilized the revised list of themes to code the entire control group and experimental group transcripts. I read and re-read the transcripts multiple times, each time better understanding the meaning of the statements shared by the participants. Upon reflection of the complete data set, I reconfigured the structure of the themes to align with the four subdomains of the appraisal inventory. Table XVI shows the final themes, definition, and examples.

**Table XVI**

FINAL THEMES, DEFINITIONS, AND EXAMPLES

| Domain | Theme | Definition | Example |
|---|---|---|---|
| Math content | Alignment | Statements referring to the state standard itself, of the degree to which the question aligns with the designated state standard. | "…that's what the standard says, right, complete the square…" |
| | Clarity | Statements referring to the ambiguity of specific text or graphics in a question. | "That it's not really clear and once you start making it clear, you're almost telling them what the answer is." |
| | Difficulty | Statements referring generally or specifically to the percent of students that will answer a question correctly, or statements concerning question aspects that impact the percent of students that will answer a question correctly. | "It's still asking them to find the difference is 430, which isn't that hard." |
| | Other | Statements containing math vocabulary, references to question content or other math problems that do not align to "Alignment", "Clarity", or "Difficulty". | "…if cosine of pi thirds equals point five, what are three other arguments that would make cosine point five?" |
| Construct-irrelevance | | Statements describing how a certain question aspect may cause a subgroup of the test taker population to perform differently due to one or more factors that are unrelated to their math ability. | "I don't think that kids understand the tax. Look like, I mean, they barely understand the tip." |
| Cognitive psychology | Multiple points of entry | Statements referring to at least two different problem-solving strategies that test takers may use to reach an answer. | "I definitely like that the kid whose thinking can do process of elimination and get the right answer. But, I don't want too much disadvantage the kid who's actually going to go through the algebra of it." |
| | Calculators | Statements containing a reference to calculators or problem-solving strategies that use calculators. | "I mean they can still plug it in." |
| | Other | Statements describing how a student or a group of students will reason through a question and/or the mathematical value of those thought processes that do not align to "Multiple points of entry" or "Calculators". | ".. would be easier than trying to memorize formulas." |
| Consequences of testing | | Statement referencing intended or unintended consequences of college admissions tests outside of the test administration. | "So, a student could get the right answer just by looking at those answers. See what I mean? Look for the commonalities." |

After rereading the transcripts multiple times and upon being satisfied with the final themes, I concluded the coding process. Appendix H presents the frequency of each code by question number and reported the summary of codes by question and group in. Overall 696 themes were assigned to the experimental group's transcript and 762 themes were assigned to the control group's transcript. Figure 6 shows the percentage of statements by theme by group for each of the four-hour educator data collection events. Both the total codes and the frequency of each theme were very similar across groups. Although the control group reviewed the first ten questions and the experimental group only reviewed the first five questions, they presented similar number of codes with similar distribution in the four-hour review session.

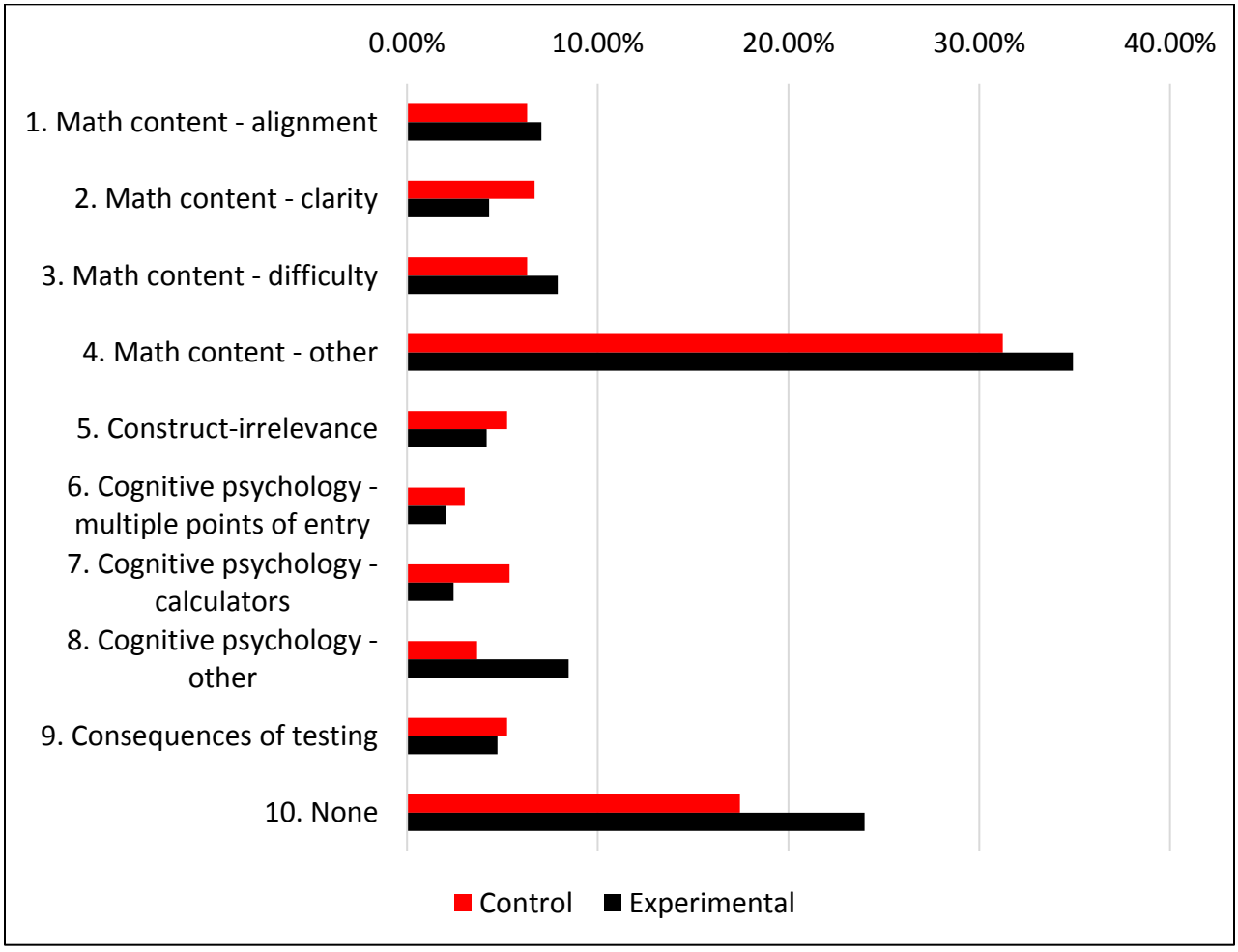



Figure 6. Percentage of themes by group – all questions.

I also compared the theme proportions across both groups for the first five questions only and the results are shown in figure 7. Figure 7 shows that the themes were distributed similarly with the largest differences being the control group had a higher proportion of statements coded to the calculator theme and the experimental group had a larger proportion of statements coded to the "difficulty" theme. There was one individual in the control group who made many comments about calculators and there was one individual in the experimental group who made many comments about question difficulty.

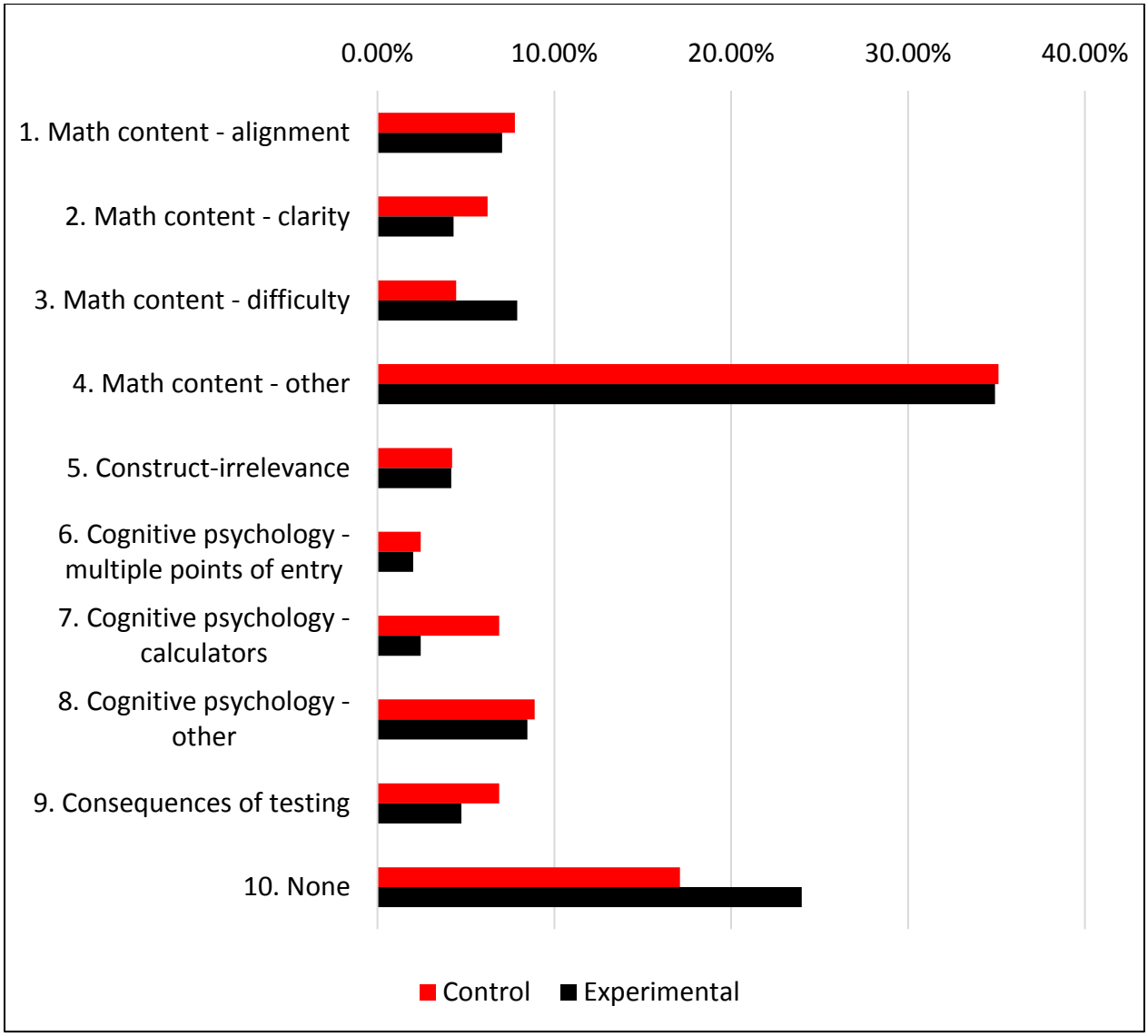



Figure 7. Percentage of themes by group – first five questions.

The final list of themes provided insight into the separate group discussions. Many of the themes were closely related to one or more of the statements on the appraisal inventory. In the cases where there was a close relationship, I investigated the pretest and posttest ratings of both groups to see whether the ratings helped me understand the discussion and recommendations. Due to the small sample size, I was not able to test whether differences were statistically significant and I limited the data analysis to descriptive statistics, which is preliminary. In the following sessions, I explore the four overarching themes in an attempt to further glean information about the quality of each test question review.

a. **Math content**

The coding process revealed the prominent themes of clarity, alignment, difficulty, and math content. In the appraisal inventory, the statements concerning clarity, alignment, and difficulty are organized under the heading of math content. Therefore, I moved the math content theme as a heading over the other three themes and included an 'other' category to classify other statements related to math content. Even though the control group reviewed twice as many questions as the experimental group, the number of control group math comments ($n$ = 385) was nearly identical to the number of experimental group math comments ($n$ = 377). In the following paragraphs, participant quotes are offset and begin with the source, or group, of the quote.

The data revealed that the control group and experimental group had very similar reactions regarding item clarity. For example, the following statements concern a lack of clarity in question 5.

> Control 7: Some of the things I had to re-read a couple times just to understand.
>
> Experimental 7: Yeah, the extra vocabulary becomes a distractor.
>
> Control 7: I was a little thrown off by the picture. I had to redraw is myself.

> Experimental 2: The original picture that's on there shouldn't be on three. Either give them a 3-D shape or don't.

Both groups ultimately revised question 5 to improve clarity and recommended to accept the question with revisions. Even though the experimental group had test taker criticism regarding clarity issues in the questions, there was virtually no difference in the specific clarity issues that both groups identified for questions 1 through 5. Table XVII shows the appraisal inventory results for the clarity statement in the sub-domain of math content. The responses to the clarity statement were similar across the two groups, and it is unclear why the experimental group's mean posttest rating was lower than their mean pretest rating.

**Table XVII**

APPRAISAL INVENTORY RESULTS FOR THE CLARITY STATEMENT

| Math content statement | Group | Pretest *M* (*SD*) | Posttest *M* (*SD*) |
|---|---|---|---|
| The question's clarity, including ambiguity or imprecision in a question. | Control | 4.13 (0.64) | 4.38 (0.52) |
| | Experimental | 4.43 (0.53) | 4.29 (0.49) |

In fact, the only noticeable difference between the two groups related to clarity was the frequency of the theme's occurrence. The control group had a total of 51 clarity references while the experimental group had a total of 30 clarity references. This difference likely resulted from the control group's review of three additional questions that were in context.

The data also revealed that the two groups had very similar reactions regarding item alignment. In question 1, both groups identified the same issue related to question alignment. The control group ultimately rejected the question because of the alignment issue, while the experimental group made edits to bring the question into alignment with the state standard. Interestingly, both groups shared their displeasure with the state alignment for question 2. They

indicated that the question was fine, but neither group thought that that standard was appropriate to emphasize on the test, as shown in one participant's reaction below.

> Control 1: It's closely aligned to the standard. Like here's how you directly assess the standard. Now, whether or not I think that's a valuable standard is a different question.

Similarly, an experimental participant suggested that the content is not emphasized in the classroom either.

> Experimental 7: I was just reading through some of the comments of the students and they're spot on. Yes, we've talked about it. But, did we spend a ton of time on it? No.

Also, both groups questioned the state alignment for question 3, but ultimately reached consensus that the alignment was appropriate. The test takers did not articulate alignment comments well for any of the questions. For example, in questions 2 and 3, they identified key words in the state standard and missed the larger intent of the mathematical concept. The experimental group had little to no advantage over the control group in this regard. The frequency of the alignment theme was very similar across groups, but not all comments were similar. For example, the experimental group frequently expressed that a given question was not assessing the entire state standard. However, the Michigan state standards were not written with the intent to be assessed with a single question. Therefore, educator judgment is required to determine if a specific question is adequately and appropriately assessing one or more parts of the state standard. The previous quote from a control group member states that question 2 is aligned appropriately, but several experimental group members disagreed during their discussion.

> Experimental 4: But even then, it's only going one way. Right, it's not doing the first part.
>
> Experimental 5: Uh, the only thing it doesn't do is test to see whether they know where the center of the circle is.

After more discussion, the experimental group reviewers better understood how to evaluate question alignment and these types of comments decreased in frequency. In addition to selected opinion differences on alignment, the groups differed in their alignment inventory ratings as shown in Table XVIII. Most notably, the mean self-efficacy increased and the standard deviation decreased for both groups from pretest to posttest, that is, both groups seemed to have a higher and less varying self-efficacy in evaluating the alignment issue after group discussion.

**Table XVIII**

APPRAISAL INVENTORY RESULTS FOR THE ALIGNMENT STATEMENT

| Math content statement | Group | Pretest *M* (*SD*) | Posttest *M* (*SD*) |
|---|---|---|---|
| The question's alignment to my state standards. | Control | 2.75 (1.49) | 3.88 (0.99) |
| | Experimental | 3.58 (0.98) | 4.43 (0.53) |

In comparison, the groups differed greatly in reactions regarding question difficulty. For example, the control group's difficulty comments were always framed in the context of all students, while on occasion, a member of the experimental group's comments explored the question difficulty for students at different ability levels as shown in the following statements.

> Experimental 7: I definitely don't think it would be easy. Medium. But again, you've got your low end, your middle, and your high end of student.
>
> Experimental 7: In my perspective I'm trying to give the majority of my class a chance at answering the question without giving it away.
>
> Experimental 7: The more I think about it, yes. Only because of like my lab class students. This is going to be a challenge to that group.

Another difference in these data was the frequency of difficulty comments was higher in the experimental group than the control group for selected items. For example, test takers indicated that questions 2 and 4 were difficult. For these two questions respectively, the experimental group had 16 and 7 difficulty comments while the control group had 5 and 4 difficulty comments. This difference suggests that test taker criticism regarding question difficulty influenced the

experimental group's discussion. Two groups had similar difficulty comments and agreed about which questions were difficult. Also, the pre- and posttest alignment inventory ratings for statement 4 under math content, "The question's difficulty, what percent of test takers will answer a question correctly" were similar across groups. Given these differences and similarities, overall the two groups were more similar than different regarding the question difficulty comments.

After consensus was reached, each participant had an opportunity to agree or disagree with the consensus on their rating form. One experimental group participant disagreed regarding alignment on question 2, one experimental group participant disagreed regarding alignment on question 3, and four control group participants thought question 6 could be revised to eliminate the math content issues if they could have spent more time on it. Other than these dissenters, participants completely agreed with the recommendations of their groups. In addition to clarity, alignment, and difficulty, participants in both groups did touch on other aspects of math content such as mathematical precision and adherence to SAT formatting style. Both groups identified math content issues, several of which were not identified by the test taker group, and resolved each issue in some way. Both groups made high quality recommendations.

b. **Construct-irrelevance**

During the initial coding of the transcripts, themes related to construct irrelevance were elusive. However, after coding the complete transcripts, I determined that some comments previously coded to the clarity theme should also be coded to a new construct-irrelevant theme. Also, several comments related to anxiety and motivation were coded to the consequences of testing theme and were also related to construct-irrelevance. After revisiting the literature, I determined that a comment concerning anxiety or motivation during test taking should be assigned the construct-irrelevance theme and a comment concerning anxiety or motivation prior to test

taking should be assigned the consequences of testing theme. In several cases, statements were assigned both themes. The number of control group construct-irrelevant comments ($n$ = 40) was larger than the number of experimental group construct-irrelevant comments ($n$ = 29). Even though the number of comments differ, the comment topics were similar and included English language learners, anxiety, and motivation, and age-related concerns. Also, many construct-irrelevant topics from the literature did not appear in either group's discussion.

Some construct-irrelevant topics were never raised, such as physical limitations, sensitivity to context, familiarity with the question format, sex or gender, race/ethnicity, geographic, socioeconomic, or religion. These topics were not discussed because the questions reviewed did not include a context that would provoke the discussion. None of the questions intentionally had issues related to these undiscussed topics, which is consistent with SAT Math Test questions.

Questions 5, 6, 7, and 10 were presented in a real-world context. The use of real world contexts often requires students to utilize less common vocabulary and may advantage students who have experiences with the context. These issues are different from question clarity. A question can be written clearly and concisely, and have construct-irrelevant issues that prevent students who know the math content from answering the question correctly.

Both groups reviewed question 5 and both groups expressed concern about the picture used to illustrate a gas tank.

> Control 1: I saw the picture and I was like this isn't a car gas tank. So, I was a little concerned about gas tank.
>
> Control 6: Kids won't understand that. We don't have propane at my house.
>
> Experimental 3: I wrote in my notes, when would a student see this object?
>
> Experimental 1: I took the shape a step further thinking through the lens of my lower EL kids. With have the shape and then actually having some of it labeled in some capacity for my low level EL kids.

Both groups went on to propose revisions to the graphic and the text to eliminate construct-irrelevant issues and both groups ultimately accepted question 5 with their revisions. The control group also identified construct-irrelevance issues in questions 6, 7, and 10. In question 6, they expressed concern that an amount of currency was referred to as a "cost" at the beginning of the question and as a "price" later in the question. In question 7, they expressed concern about the quantity of words in the question, and that many of the words were not needed to solve the problem. In question 10, the topic was profit, and when the profit was negative, the value was presented in parentheses, "($285) loss". This style is used in accounting, and the participants preferred to present the value as a negative number. It is unclear whether the experimental group would have identified the same issues in questions 6, 7, and 10. The test takers rarely connected question issues with construct-irrelevance. When they were confused or distracted by part of a question, they often indicated that problem was the question difficulty and not construct-irrelevance. The test taker's criticism for questions 6, 7, and 10 were unlikely to have helped the experimental group identify construct-irrelevant issues. However, based on the issues they identified in question 5, and based on the criticism that the test takers shared, it is likely that the experimental group would have discussed and resolved construct-irrelevant issues for these questions.

The test taker group, the control group, and the experimental group all reported that fractions cause anxiety in test takers.

Test Taker 1: Nobody likes fractions

While revising a test question with the test takers, I inserted a fraction and test taker 2 said: "Don't make it a fraction. Because fractions just make everybody…" While discussing question 3 with the control group, fraction anxiety was a topic.

> Control 1: I think [participant]'s point about anxiety though is real too. I mean fractions give kids anxiety. I mean especially if it's on the no calculator. If it's on the calculator, a little less anxiety.
>
> Control 6: They just shut down.
>
> Control 2: Or if they get a fraction, then they think it is wrong.

A similar comment concerning question 3 was made by the experimental group.

> Experimental 7: I would say the medium difficulty ranking may come from fear of fractions.

Even though there was consensus that fractions cause anxiety, there was consensus by all three groups that fractions should appear in some test questions. On the appraisal inventory, the fifth statement under construct irrelevance was related to anxiety, and the group results are shown in Table XIX. The experimental group's mean rating from pretest to posttest increased more than the control group's mean rating despite the similarity in their conversations.

**Table XIX**

APPRAISAL INVENTORY RESULTS FOR THE ANXIETY STATEMENT

| Construct-irrelevance statement | Group | Pretest *M* (*SD*) | Posttest *M* (*SD*) |
|---|---|---|---|
| There is unfairness due to a construct impacting test anxiety. | Control | 3.50 (1.20) | 4.00 (0.53) |
| | Experimental | 3.29 (0.76) | 4.43 (0.53) |

The appraisal inventory statement related to motivation showed similar results, as shown in Table XX. These data are consistent with the group conversations in that we discussed and resolved some motivation-related issues in both groups.

**Table XX**

APPRAISAL INVENTORY RESULTS FOR THE MOTIVATION STATEMENT

| Construct-irrelevance statement | Group | Pretest *M* (*SD*) | Posttest *M* (*SD*) |
|---|---|---|---|
| There is unfairness due to a construct impacting test taker motivation. | Control | 3.38 (1.06) | 3.75 (0.89) |
| | Experimental | 3.14 (0.90) | 4.43 (0.53) |

After the control group completed their discussion about fractions in question 3, a participant asked the following question about motivation, which was met with an affirmative response from other participants.

> Control 3: Would they maybe even just look at the item and see the answers with a bunch of fractions and then walk away from it at that point?

Motivation comments were not always connected to fractions. For example, during the discussion of question 4, a control group participant was concerned that the difficulty of the question would cause students to give up and the experimental group expressed concern that the difficulty would reduce motivation on subsequent questions. The test taker group did not have comments related to motivation on question 3, but one participant made the following comment about question 4, which shows how difficulty and motivation are related.

> Test taker 6: And so, like if I was taking the SAT and I saw this question I would just, honestly, I would just have given up and either guessed or waited until the end and like whatever time I had left I would have tried to solve it.

All three groups ultimately accepted question 4 with revisions to reduce anxiety, and difficulty.

A small number of construct-irrelevant comments are not related to context, anxiety, or motivation. For example, participants in both groups made a small number of comments concerning how a test taker's age may impact their understanding of question. Also, one comment in each of the groups was about a question being 'frustrating.' Like the math content comments, both groups identified and resolved construct-irrelevant issues appropriately. The approved questions from both groups lack construct-irrelevant issues. Question 6 was ultimately rejected by the control group mainly due to construct-irrelevant issues. Several control group participants disagreed and thought the question could be revised appropriately if more time was spent on it. However, due to time constraints, rejecting the question was an appropriate decision.

The mean appraisal inventory ratings for all construct-irrelevance statements suggest that the reviewing test taker criticism resulted in a much larger increase in educator self-efficacy for the experimental group than the control group, as shown in Table XXI.

**Table XXI**

APPRAISAL INVENTORY RESULTS FOR THE CONSTRUCT-IRRELEVANCE DOMAIN

| Statement | Group | Pretest *M* (*SD*) | Posttest *M* (*SD*) |
|---|---|---|---|
| Questions may be asked in a way that is unfair to one or more subgroups of test takers. Rate your degree of confidence in evaluating a question's unfairness to test takers as a result of each of the following statements. | Control | 3.73 (0.38) | 3.88 (0.33) |
| | Experimental | 3.66 (0.52) | 4.35 (0.63) |

c. **Cognitive psychology**

Initial cognitive psychology themes that emerged during coding included multiple points of entry and calculators. The 'multiple points of entry' theme refers to the number of available problem-solving methods by which test takers may solve a problem. The calculator theme is also connected to problem solving. The SAT has a no calculator portion and a calculator portion. If a question is on the no calculator portion of the SAT, then the question reviewers know that test takers must solve the question without a calculator. If a question is on the calculator portion, then educators hypothesize what additional problem-solving strategies are available to test takers who chose to bring and use a calculator.

In addition to multiple points of entry and calculators, other cognitive psychology comments were shared, but the comments were not frequent enough to warrant separate themes. For example, participants discussed whether the question type (4MC or SPR) influenced the way test takers solved the problem. Also, participants discussed what previously memorized formulas

were needed to solve problems. I created a cognitive psychology overarching theme to include multiple points of entry, calculators, and other statements related to cognitive psychology. The number of control group cognitive psychology comments ($n$ = 92) was nearly identical to the number of experimental group math comments ($n$ = 90).

Both the control group and the experimental group had multiple comments regarding multiple points of entry in questions 1 and 3. In question 1, both groups identified the exact same approaches to solve the problem: (a) recall memorized trigonometry identities (b) draw or imagine a unit circle and (c) draw or imagine a 30-60-90 triangle. The same was true for question 3, and the approaches are referenced in the following statements.

> Control 2: I thought if they have their calculator and they have multiple choice, then all they have to do is plug those in and use the solver on their calculator and see which one gives them the answer.
>
> Experimental 2: I think their initial instinct would be to go to the quadratic formula.
>
> Control 1: If they know the quadratic formula and they know that the thing under the square root has to equal zero in order to get one solution, they'll just set that equal to zero and solve it.
>
> Experimental 7: That question I think, makes them use the discriminant, where on this I was able to use completing the square.

Feedback from the test takers indicated that they were familiar with this type of question, but they did not elaborate on how they solved it or how they would solve it. These comments did not help the experimental group identify issues related to multiple points of entry. The control group also shared comments related to the multiple points of entry theme on questions 8, 9, and 10, but the experimental group did not discuss those questions. The test taker criticism did not include any comments regarding multiple points of entry on questions 8, 9, or 10. Based on the similarities between the groups for questions 1 and 3, I expect that the experimental group would have discussed multiple points of entry on questions 8, 9, and 10 even without test taker criticism. The

appraisal inventory ratings were similar for both groups on the statement referring to weak and strong problem-solving strategies. As shown in the participant's statements, cognitive psychology themes intermingle more than math content themes and it was common for a statement to be coded to both multiple points of entry and calculators.

Of the questions reviewed by both groups, questions 3, 4, and 5 were designated for the calculator portion and the other questions were designated for the no calculator portion. Participants could recommend a change to the test portion, and both groups did discuss the possibility on certain questions.

> Experimental 6: And then, the calculator requirement. Does this question, should it be, I mean it doesn't need a calculator.
>
> Control 3: I think that if we really wanted to salvage it, for something, we might bump it to the no calculator and put in some of the more standard forms for the solution.

In the final recommendations, the control group recommended to move questions 3 and 4 from the calculator portion to the no calculator portion. In the control group, one participant discussed whether certain types of calculators provided an advantage over other, less powerful, calculators.

> Control 3: My question is, because I don't use all the different types of calculators. But, on the calculator exam aren't there different types of calculators where you can say define f(x) for this? What is f(x) and then it will pop out. Am I correct?
>
> Control 3: I thought very similar but I was also thinking if they have a CAS, computer algebra system, then they just type in the equation and say solve and it pops up and gives them the answer.

The experimental group did not make any comments related to the advantages of more powerful calculators. Other than this discrepancy, the comments that both groups shared related to calculators were similar.

When discussing questions 2 and 5, both groups considered the impact of the question type, 4MC or SPR, on the way the student solves the problem. On question 2, the correct answer was the

square root of 21. A common student misconception leads to the correct answer of 21. However, 21 was not presented as one of the answer choices and both groups agreed that, in this case, the inclusion or exclusion of 21 as an incorrect answer would have a large impact on the question performance. Ultimately both groups chose not to include 21 as an incorrect answer on question 2 because it weakened the question alignment. On question 5, at least one person from each group recommended that the question be changed from SPR to 4MC.

> Control 3: I don't. I mean, I think this should be maybe multiple choice. I don't know. Put that as a SPR, I can't answer.
>
> Experimental 5: It seems like this one would be better to have four choices. Because then if they have rounded incorrectly a little bit, they're gonna see which one is close to their answer.

The control group recommended to change the question type to 4MC and the experimental group recommended to keep the question type as SPR. In addition to having similar comments regarding question type, the control group had similar results to the experimental group on the appraisal inventory statement related to question type, as shown in Table XXII.

**Table XXII**

APPRAISAL INVENTORY RESULTS FOR THE QUESTION TYPE STATEMENT

| Cognitive Psychology statement | Group | Pretest *M* (*SD*) | Posttest *M* (*SD*) |
|---|---|---|---|
| The question's response format (MC or SPR) provides insight into the student's thought process(es). | Control | 3.50 (0.76) | 4.00 (0.76) |
| | Experimental | 3.43 (1.27) | 4.29 (0.76) |

A formula sheet is provided during the SAT administration, but some test questions need formulas that are not listed on the formula sheet. For example, question 1 could be solved using one or more memorized trigonometric identities and question 2 could be solved if test takers remembered the standard form of the equation of a circle. Both groups made comments about memorizing content and they also commented that students may need to remember the quadratic formula to solve

question 3. On the appraisal inventory statement for long term memory, the experimental group's mean posttest rating increased more than the control group, as shown in Table XXIII. However, overall the group discussions were very similar in terms of content and quality.

**Table XXIII**

APPRAISAL INVENTORY RESULTS FOR THE LONG-TERM MEMORY STATEMENT

| Cognitive Psychology statement | Group | Pretest *M* (*SD*) | Posttest *M* (*SD*) |
|---|---|---|---|
| An appropriate amount of long-term memory was required in order to answer the question. | Control | 3.50 (0.76) | 3.75 (0.71) |
| | Experimental | 3.57 (0.79) | 4.29 (0.76) |

One aspect of cognitive psychology, question point value, did not directly emerge in the comments of either group. I did not locate any comment which addressed whether the cognitive processes required to answer the question correctly warrant one additional raw score point. This aspect was eluded when participants discussed the impact of calculators or multiple points of entry, but question point value was not directly commented on. Therefore, the comments and the appraisal inventory results for both groups were very similar under the theme of cognitive psychology. The final versions of the questions approved by both groups are appropriate in terms of cognitive psychology. Lastly, all the participants from both groups agreed with the consensus ratings regarding the issues related to cognitive psychology.

d. **Consequences of testing**

I classified several comments from both the control group and the experimental groups ($n_{control} = 40$, $n_{experimental} = 33$) under the consequences of testing theme. The frequency of the comment topics did not warrant the creation of subthemes. Also, there was a noticeable absence of certain topics related to consequences of testing. The differences in comment topics across the control group and experimental group were more pronounced than in the other

appraisal inventory domains and there were several similarities across the group's comments as well.

Both groups discussed opportunity to learn, which is related to consequences of testing because test content may influence when students are taught certain content. The comments in both groups were often from the high school teachers as they described when math content was taught. Concerning the state standard assessed by question 2, high school participants in both groups reflected on their own teaching experiences.

> Control 6: One section in geometry. When I taught geometry, it was like one day, barely a fit.
>
> Experimental 7: They've seen it for 3 days in geometry at some point. I was just reading through the comments of the students and they're spot on. Yes, we've talked about it, did we spend a ton of time on it? No. We do completing the square in Algebra 2, and Geometry. And Algebra 1.

Also, one of the higher education participants from the control group added their thoughts.

> Control 2: I've taught college for four years and I've never once needed this from my students. They don't remember how to do it and they can graph it so it's not important.

Participants also discussed opportunity to learn in terms of accelerated content exposure. A high school participant from the experimental group expressed concern that some students may have taken geometry as eighth graders or high school freshmen and therefore the content in question 2 would not be fresh in their mind. This comment refers to students who are on an accelerated path. The earlier participant comments refer to students who are on a non-accelerated path or a less-accelerated path. Therefore, the conversations of both groups excluded students who are not exposed to this content prior to taking the SAT. From the consequences of testing perspective, we could have discussed whether certain subgroups of students (gender, race/ethnicity, first language, socio-economic status, special education, etc…) are disproportionately represented in the

accelerated path. As I noted earlier, compared with the recommendations in existing literature, these topics are noticeably absent from the comments.

Both groups also discussed the *xy*-plane, fraction anxiety, and the use of π in question 5. Question 2 begins, "In the *xy*-plane, what is…" and the participants said that the use of "In the *xy*-plane" is unique to the SAT, but they are changing their teaching to accommodate.

> Control 7: I just want to make a vocab comment, again coming from an algebra class, and we talked about this this summer. But, this *xy*-plane. I mean it's not brought up ever. Like we don't. I've never used it until last year. We always, like, the coordinate plane. If they're going to use *xy*-plane then I think that's something in my district that our middle school should start using the *xy*-plane. Cause we know it when we were looking at practice questions to give our kids, we say *xy*-plane, *xy*-plane. So, our kids didn't know what *xy*-plane was. Like we're going to now use *xy*-plane all the time.
>
> Experimental 2: The only problem, uh, the *xy*-plane. I don't. That's just. We've started calling it the *xy*-plane just because we know it shows up on the SAT, to get them familiar with that. I've never called it the *xy*-plane before so I think some kids would be confused by that. You know. If it's only been presented to them as the coordinate plane, you know they might be like what's the *xy*-plane?

This impact to classroom teaching is an unintended consequence. The control group recommended to accept question 2 as is, including "In the *xy*-plane,". The experimental group recommended to accept question with revisions, and their revised question did not reference the *xy*-plane, which agrees with the test taker's recommendation. In addition to the *xy*-plane, participants in both groups also discussed the anxiety that test takers have when fractions are present in a question and when test takers obtain a fraction for their answer. Many of the comments made were in the context of a specific question, but applied globally to all questions with fractions, and fraction anxiety is an unintended consequence and neither group eliminated fractions from the questions. Lastly, on question 5, participants from both groups discussed the use of π in the question solution strategy. To answer question 5, test takers may use an approximation of π in their calculations. The participants in both groups discussed whether this test question was promoting the use of 3.14 or π.

> Control 7: π or 3.14? I mean a lot of the test, do we use π or do we use 3.14? I did them both and I got different answers. So, I'm like this question specifically, a lot of kids on the geometry test here we tell them to use 3.14. So, I think the teacher has to be on the same page as the teacher or the SAT.
>
> Control 6: Then we should still be able to use both [π and 3.14] in our classrooms as far as changing teaching.
>
> Experimental 1: I feel like students that, like in Algebra 2, we use π all the time on our calculator. Like we deter them from using 3.14.

Promoting the use of 3.14 rather than π in the classroom is an unintended consequence. The control group revised the question so that students could use either π or 3.14 to reach the correct answer. The experimental group also revised the question so that students could use either π or 3.14 to reach the correct answer, and they also provided the note, "(Use π = 3.14)" in the question. In each of these three cases, *xy*-plane, fractions, and π, the identical issue was raised in both groups and resolved. Like the opportunity to learn discussions, these comments lacked further depth into whether certain subgroups of students are impacted by the issues. These data do not point to one or more obvious subgroup(s) that may be disproportionately disadvantaged by the recommendations made.

The comments regarding the use of π or 3.14 by both groups were set in the context that all students had a choice to use the exact value of π on their calculator or the rounded version 3.14. The control group did not discuss whether all students had access to a calculator. A participant from the experimental group asked whether the students at the participating high school all had calculator access during the SAT and one of the high school teachers confirmed that the school provides calculators to all students for the SAT. Another difference between the control group and the experimental group is that the experimental group discussed test preparation strategies and the control group did not. A participant from the experimental group described a strategy for identifying the correct answer in a multiple-choice question by identifying commonalities between

the answer choices. Then, using question 4 the participant showed how to implement the strategy. Also, different experimental group participant described one way they prepare students for the Preliminary SAT (PSAT) which is administered during freshman year.

> Experimental 2: And we run into that when we practice for the PSAT at the ninth grade level. We specifically do non-calculator stuff and calculator stuff.

Teaching students to utilize non-content based strategies for identifying correct answers is an unintended consequence, but practicing math with and without calculators is a desired consequence which is also encouraged by post-secondary math educators.

There are two more outlier statements of interest related to consequences of testing. One comment is from a control group participant concerning question 8. Question 8 is SPR format and there is more than one possible answer that students can enter and receive credit.

> Control 1: I like that it's open ended and that there's more than one correct answer. I think that when it comes to standardized tests, this is an evolution of standardized tests that, like philosophically, I believe in. And I think will affect the way that teachers teach. I think teachers would teach differently towards this question than they would towards your more traditional multiple choice standardized test question.

The second comment is from an experimental group participant concerning question 1. In question 1, students are provided with an equation in one variable and asked to identify another equation in one variable that has the same solution.

> Experimental 4: But I can imagine them, again what I know about students is they're often trying, they see mathematics is about getting a solution, not necessarily about seeing a relationship.

These two comments are similar in that they reflect changes to the math curriculum that are inspired by the state standards, not by the SAT. This consequence of changes to the state standards is intended, and is an intended consequence of the SAT question.

The revised questions that the control group and the experimental group approved are likely free from issues related to consequences of testing. There are aspects of consequences of

testing that were not discussed and only the experimental group discussed test preparation strategies and calculator availability. Other than these differences, there were many similarities between the group discussions.

In this section, I summarized the results of the triangulation analysis of the quantitative and qualitative data available. The summary was presented in the framework of the four appraisal inventory domains. Within each domain, I compared the control group results with the experimental group results. During data collection, the two groups differed on two aspects. First, the experimental group included seven persons instead of eight and second, the control group finished discussing the first 10 test questions while the experimental group only finished discussing the first five test questions in the given four hours. However, the quantitative and qualitative data available showed great similarity between the two groups. In the discussion sections, I will further elaborate on the limitations of the results.

## C. **Summary**

Data collected for this experimental study revealed much about educator self-efficacy and test review quality. I collected test taker criticism on a set of test questions and summarized their feedback. I used ANCOVA to investigate the impact of test taker criticism on educator self-efficacy. I used educator rating data, transcriptions of educator question reviews, and written feedback from educators to evaluate the impact of test taker feedback on test question review. From these data, I derived themes and summarized the themes with descriptions that were tied tightly to the literature base. In the next chapter, I will draw conclusions based on multiple sources of evidence and discuss about the limitations of the study.

# V. DISCUSSION

This experimental study of math educators showed both similarities and differences between the control and the experimental groups. Both groups significantly improved their test question review self-efficacy. With the test taker criticism available, the experimental group improved their self-efficacy more than the control group. The question review process yielded high quality question recommendations from both groups. The qualitative analyses revealed only subtle differences in question review quality between the two groups. I discuss the results in more detail below.

## A. **Research Question 1**

Nearly all participants had no prior experience with reviewing standardized test questions. Split-plot ANOVA showed that both groups scored significantly higher on their overall educator self-efficacy posttest. ANCOVA showed the experimental group reported significantly higher educator self-efficacy than the control group, after controlling for their pretest total score. Preliminary investigations into the four educator self-efficacy domains revealed that (a) both groups improved significantly from pretest to posttest on each of the domains; (b) after controlling for their corresponding pretest total score, the experimental group scored significantly higher than the control group on three domains, including construct irrelevance, cognitive psychology, and consequence of test, but not on math content.

Following the theory in Bandura (1997), I classify the participation in the question review process as a mastery experience, and the experimental group's review of test-taker criticism as a vicarious experience.

Previous findings suggest that participation in a mastery experience results in an increase in teacher self-efficacy (Morris et al., 2016). In this study I utilized an instrument to measure

educator self-efficacy in question review and explored a mastery experience to improve educator self-efficacy in question review. These findings suggest that educator self-efficacy increased in both groups after the participants completed the mastery experience and this result is consistent with previous research on teacher self-efficacy.

A vicarious experience is also likely to increase teacher self-efficacy, especially if the model is perceived as similar to the participant (Morris et al., 2016). In this study, the model was not similar to the participant (i.e., student versus teacher), but Morris et al. (2016) also the suggests that a vicarious experience will likely increase teacher self-efficacy if the model struggles to overcome obstacles. The experimental group observed many occurrences in which the test takers struggled to identify and resolve issues. I suspect that being able to see test takers' criticism on the questions helped the experimental group educators to be more confident in their question review than the control group educators who did not have access to such information. The significantly larger increase in educator self-efficacy for the experimental group is consistent with my pilot study (Trapp, 2015), and is consistent with previous research on teacher self-efficacy.

B. **Research Question 2**

The question review quality analyses found more similarities than differences across groups. To examine question review quality, I utilized coded transcriptions, responses to the open-ended questions on the rating form, initial question ratings, final question recommendations, and appraisal inventory responses. A single source of evidence is insufficient to evaluate question review quality, therefore I used triangulation analysis which involved these quantitative and qualitative data.

I utilized lessons learned in other fields to inform my data collection procedures. For example, psychological test developers have recommendations for conducting question reviews,

and many recommendations are applicable for math tests (Vogt et al., 2004). On the other hand, the results of this study will likely have applications in other fields that use subject matter experts to evaluate standardized test questions. The main differences between this study and other existing literature is that this study uses an experimental design and mixed methods for data collection and analysis and neither was present in other studies that I reviewed.

Prior research warns that undesirable rating effects may appear during educator question reviews. K. E. Ryan (2002) expressed concern about confirmationist bias in that educators may help identify the strengths of the test questions that make them appropriate for the purpose of the SAT without looking carefully enough at the weakness of the test questions that make them less appropriate. This is the first study that uses quantitative and qualitative data to investigate the phenomenon of educator question reviews and therefore provides a new perspective on the presence or absence of confirmationist bias. The data in this study shows little confirmationist bias in the areas of construct-irrelevance and consequences of testing. The participants could have evaluated the questions more carefully from the perspective of test takers at more ability levels and from the perspective of test takers who do not follow a normal or accelerated math course path. However, given that the majority of the participants had never participated in this process before and that these are difficult question aspects to evaluate, the participants in both groups should be commended for the many times that they identified and resolved issues related to construct-irrelevance and consequences of testing in the questions reviewed.

The success of the groups in avoiding confirmationist bias was likely due to the diversity of each group rather than the training provided. I intentionally ensured that each group consisted of high school and post-secondary math educators. Also, within each group there was an attempt to have diversity in gender, race/ethnicity, math area of expertise, years of teaching experience, and

course teaching experience. Therefore, questions were reviewed from many perspectives in each group and some participants felt the diversity of voices was appropriate.

> Experimental 1: I believe we were very thorough in examining the questions from many perspectives. I also think that everyone from the team brought a unique set of skills and points of view that added to the quality of the questions we built together.
>
> Control 8: We had a diverse group of educators giving suggestions. Everyone has different backgrounds as far as teaching and were able to use those experiences to enhance questions.
>
> Control 2: I think there are enough people present from varied secondary/college backgrounds that a variety of viewpoints are present.

While one participant felt even more voices would be better.

> Experimental 5: Maybe including a more higher ed participants to review the questions, since high school courses are meant to prepare for higher ed.

These qualitative data suggest that this study avoided confirmationist bias to the greatest degree possible. Lastly, not only were the groups diverse, but all participants contributed without needing to be called on. Like all groups, some participants are more vocal than others, but no voices went unheard.

The initial question ratings from the participants yielded low interclass coefficients for both groups. The poor interrater agreement for each group are expected and confirm one of the reasons why multiple raters participate in the process. That is, different participants identify different question issues and assign different levels of severity to those issues. The participants reached consensus and generated group recommendations after group discussion. In this case, the variety of their initial recommendations indicates that the participants in each group were diverse and having one fewer participant in the experimental group did not appear to limit the diversity of opinions. Based on these results, it is likely that the experimental group would not have improved their low initial interrater reliability had they reviewed more questions.

Following the initial ratings, the groups discussed the quality of the questions. The transcripts revealed strong similarities in the content and quantity of the themes. The themes that emerged were closely tied to the literature and the topics found in the appraisal inventory. These similarities exist even though the experimental group reviewed five fewer questions than the control group. I compared the proportion of each theme's appearance in the first five questions for both groups and found that the experimental group had a higher proportion of math content themes. This finding agrees with the final recommendations in that the experimental group took more time to revise and ultimately accept questions than the control group. Aside from the higher proportion of math content themes in the experimental group, the proportion of comments for the other themes was similar.

Both groups identified several issues regarding math content, including comments related to clarity. Wynd and Schaefer (2002) suggested that an expert review panel should identify problems with wording, clarity of meaning, and construct of the item in terms of English grammar on a nursing assessment and this was also an expectation of the educators in this study. The groups did identify and resolve several clarity issues. Also, when the groups chose to revise a question, they evaluated the result, to ensure that all questions are clear, unambiguous, and grammatically consistent.

In a study about educators' ability to predict student performance, Coladarci (1986) found that the educators accurately judged question difficulty, but their predictions for high performing students were more accurate than their predictions for low performing students. These results were not replicated in this study because neither group was successful at evaluating a question's difficulty from the perspective of different student ability levels. The test taker criticism included comments about question difficulty, but their comments were often entangled with construct

irrelevance and consequences of testing issues rather than difficulty for students based on their understanding of the math content. Several questions were difficult for all test takers and prevented them from discussing differentiated ability levels. The experimental group had minimal comments on differentiated difficulty, but took little to no action. The control group did not deviate from general difficulty level in their comments. Therefore, to some degree, these data are consistent with Coladarci's findings in that neither group gave useful insight into the question difficulty for low performing students.

I was unable to locate prior studies that investigated educator ability to identify construct irrelevance issues in questions. Although the experimental group's self-efficacy increased in the posttest ratings, they did not identify and/or resolve more construct-irrelevant issues than the control group. The control group identified several construct-irrelevant issues in questions 6, 7, and 10. When reviewing test questions, the test takers often misidentified construct irrelevant issues as question difficulty. Therefore, their comments on questions 6 through 16 would not have provided an advantage to the experimental group over the control group.

Pellegrino et al. (2001) suggested assessment developers pay attention to three elements of the assessment triangle--cognition, observation, and interpretation-- and their coordination. Participants from each group made multiple comments regarding cognition including themes such as multiple points of entry, the impact of calculators on problem solving, and memory. There were several calculator-related statements from both groups and the control group specifically had unique statements regarding the changes to problem solving that occur when advanced calculator functionality is available. Participants in both groups successfully recognized the difference between experts and novice and they suggested that 'experts' will have an advantage over less-prepared test takers on questions 1 and 3. Pellegrino et al. (2001) explains that experts, when

presented with multiple problem solving strategies, use metacognitive skills to evaluate which strategy to implement and monitor the success of implementing the strategy. The participants in both groups commented on multiple points of entry to solve problems, but they did not make the connection to metacognition of the possible strategies as Pellegrino et al. proposed. These participants may develop expertise in evaluating questions from the perspective of metacognition if they participate in additional test question reviews.

Participants are also likely to develop their question review skills regarding consequences of testing with more practice. Statements classified under the consequences of testing theme refer to both intended and unintended test consequences. Concerning the SAT, an example of an intended consequence is the appropriate use of SAT test scores in college admission decisions. An example of an unintended consequence is the use of SAT test scores for job eligibility (Zwick, 2006). Standardized test score misuse has occurred and the consequences of those misuses have different levels of severity for different persons. K. E. Ryan (2002) uses the term 'disproportionate consequences' and Kane (2013) uses the term 'differential impacts on groups' to call attention to such unintended consequences. Both groups had mixed success in discussing disproportionate consequences of the questions. Comments concerning opportunity to learn remained relatively superficial, which is somewhat expected because the focus of the task is question evaluation not curriculum evaluation. The same could be said for the comments regarding test preparation practices and other impacts to instruction. Time spent delving into those topics reduces the number of questions that can be reviewed, and both tasks must be balanced appropriately. Given the enormous numbers of issues related to consequences of testing, the group of mostly inexperienced question review participants did well, but neither group outperformed the other.

The two groups had similar test question review quality and neither excelled at any one aspect. As noted earlier, the test takers missed several issues in questions 6 through 16. If the experimental group had reviewed more questions, I do not expect that the test taker comments would have provided any advantage over the control group and I would have likely found that neither group generated significantly more high-quality recommendations than the other group.

There were many similarities between the groups, but evidence suggests that the test takers influenced the experimental group participants. For example, the difference in the number of questions reviewed may be the result of the experimental group's commitment to resolving the test taker's concerns. The experimental group took the time to resolve all issues in the questions they reviewed, even when I suggested to reject a question and move on. Also, the experimental group's final recommendations on questions 2, 3, 4, and 5 were clearly influenced by the test taker's comments. Many of the test taker's recommendations were implemented, along with other adjustments by the experimental group. Additionally, the appraisal inventory results suggest that a portion of the experimental group's educator self-efficacy increase was due to the exposure to test taker criticism. I suspect that the increase in educator self-efficacy resulted in confidence and eagerness to resolve the issues raised by the test takers. However, being influenced by test taker criticism did not result in higher question review quality in this study.

This experimental study revealed that the exposure to test taker criticism had little to no impact on question review quality. I suspect the reasons might be that the educators are more knowledgeable about the math content and more experienced with the test than test takers and working closely with students in teaching allow them to evaluate the questions from test taker's perspectives as well. In particular, the educator participants in this study had at least two years of teaching experience, which allowed them to have a good knowledge about the math content, the

SAT test, and the students. Although the control group reviewed more questions than the experimental group, the data suggests that the experimental group would not have been any more or less successful in their discussions or final recommendations had they reviewed more questions. The theoretical base for evaluating educator self-efficacy and question review quality permitted a rigorous, multi-faceted evaluation of the variables. This study, being the first to utilize mixed-methods to study educator question reviews, also provided an opportunity for students to take part in the question review process. Although the findings suggest that test taker criticism did not significantly influence question review quality, this does not suggest that test taker comments are not worth collecting. The results of this study agree with the findings of Vogt et al. (2004) in that test takers can enhance the validity evidence of the test, if they are consulted early enough in the process. The validity evidence of the test is enhanced when educator self-efficacy of participants is maximized.

## C. **Limitations of the Study**

First, the data collection location and participants are constrained. I collected quantitative and qualitative data for this study from students and math educators who reside in a portion of Michigan due to the limit in time and cost. The purpose of this study was to investigate the impact of test taker question criticism on the standardized test question review process. Because the SAT is administered domestically and internationally, a more representative sample of test stakeholders would have included participants from more locations. I selected this location in Michigan because of my existing professional relationships with two teachers at the local high school and because the population of test takers and math educators had desired demographic diversity. The data collection site was cost-effective and was unfamiliar enough that I didn't feel 'at home' and helped create an objective circumstance. However, limiting the population of potential participants to

those in or around a small city in Michigan made the data collection events a localized educator question review and it is unknown whether the same results would have been obtained using participants selected from the national population, which is common for a standardized test review process.

Second, a limited number of validation tests were completed for the appraisal inventory in this study. This study developed appraisal inventory for educators to self-report their question review self-efficacy. The inventory was designed using practices recommended by Bandura (2006) and was piloted and further refined. Due to the small sample size, I could only complete a limited analysis of the inventory data collected for this study and examine the internal consistency of the inventory. Although all analyses suggest that the analyses conducted were appropriate, further evidence is desired to ensure that interpretations from the data are defensible.

Third, the two groups reviewed a different number of test questions and the total number of questions reviewed was small. The question set used for this study included a sample of the content assessed on a 58-question SAT Math Test, and neither group of educators reviewed enough questions to be considered representative of the test. Therefore, the results of this study may not generalize to the entire content breadth and depth assessed on the SAT Math Test. The question set included questions with and without issues and each question generated discussion in each group. Of the 20 questions, the test takers reviewed 16, the control group reviewed ten, and the experimental group reviewed five. Each group had a rich discussion and the resulting transcripts had nearly equal length. In general, the experimental group worked on revising a question until they could approve it while the control group moved on more quickly rather than spending time on question revisions. Therefore, I had fewer question results to compare across the groups than desired. The triangulation analyses of the quantitative and qualitative data yielded

results that would have likely sustained even if both groups reviewed the same number of questions. However, the conclusions are still based on a limited amount of data and can be further examined with a larger test question sample.

Finally, the sample size for both educators and students is small. In practice, test developers include eight participants in question reviews to achieve demographic diversity. To make my study consistent with common practice in the field, I recruited eight educators for each group. The number of participants in this study, however, prevented me from conducting data analyses that require a larger sample size. For example, I could not investigate the relationship between educator demographics (e.g., years of teaching experience) and educator self-efficacy. Also, preliminary analyses at the domain level of the self-efficacy inventory revealed some assumption violations. The assumption violations and small sample size did not allow me to make decisive conclusions about the impact of test taker criticism on the educator self-efficacy domains. In addition, the small number of students and lack of SAT performance information about the students constrained my ability of examine the nuance of students' contribution in this process. For example, with the current data available I could not tell what types of students may be more likely to provide information that can be helpful for educators than others.

## D. **Directions for Future Research**

The limitations of this study present opportunities for future research. Researchers could attempt to replicate these findings with participants from other locations, with more educators, more students, and more test items. The qualitative data analysis utilized a type of coding that was summarized in terms of theme frequency, but the use of other types of coding such as magnitude coding or theoretical coding (Saldaña, 2009) in future research may provide further insight in the impact of test takers' input on educators. This study could also be replicated on question reviews for different content areas (e.g. English/language arts, science, social studies) or a different type of content completely (e.g. nursing). This study benefitted from research conducted in the nursing field by Berk (1990), Grant and Davis (1997), and others, and I believe the results of this study and replications of this study will benefit the development of other types of standardized tests.

The appraisal inventory developed for this study captured self-efficacy ratings from teachers regarding a specific task – standardized test question review. Further validation tests of the inventory could be completed using a larger sample size. Industry experts could review the inventory statements and judge statement relevance. Ideally, a large enough sample could respond to the survey such that each response category (highly unsure, unsure, neutral, confident, highly confident) for each statement receives at least 10 responses. The resulting data could be analyzed with a Rasch model. Analyses might include a review of the item and person fit statistics, a rating scale analysis, and a dimensionality analysis (Wolfe & Smith, 2007). If data from other measures are available, then a multitrait-multimethod analysis will also provide useful data for evaluation. Lastly, the dimensionality of the scale could be evaluated with a factor analysis.

Some researchers have reported that ratings scales which utilize a neutral category in the response options may not yield the desired statistical properties (Bradley, Peabody, Akers, &

Knutson, 2015; DeMars & Erwin, 2004). For example, each response option should be the most frequently selected option for a range of logits on the measurement scale. An analysis of an appraisal inventory administration with an adequate sample size may reveal that "neutral" is never the most frequently selected option for any logit range. Therefore, future research could investigate whether the neutral category in this appraisal inventory should be used according to desired statistical properties.

Another area of future research is the use of test question review as an instructional method. The test taker criticism collected revealed that students evaluate questions differently than educators do. They describe their confusion, motivation, confidence, anxiety, as well as their content knowledge when they evaluate questions. On several occasions, test takers explained their successful and unsuccessful problem-solving strategies to the rest of their peers. The purpose of assessment is to learn what students have learned and what they have not mastered, therefore reviewing test questions can provide an opportunity for teachers and students to better understand student learning.

Future research can explore other review formats as well. I utilized one format of educator question review in this study and there are several other less expensive and less disruptive ways to facilitate question reviews. Other review formats may allow educators who are often quiet in a group setting to share more thoughts than they would have been. For example, one possible format allows participants to provide their comments electronically and participants may be able to view and respond to question comments made by other participants. Also, other review formats may be more conducive to eliciting more comments related to construct-irrelevance and consequences of testing. From a test stakeholder perspective, methods that results in higher the educator self-efficacy and higher question review quality are preferred over other methods.

## E. **Summary**

The purpose of this study was to investigate the impact of test taker question criticism on educator self-efficacy and question review quality during the standardized test question review process. I utilized an experimental design and provided the experimental group with test taker criticism. I collected pre- and posttest data concerning educator self-efficacy and question review quality data through consensus recommendations, rating forms, group discussions, and a final survey. The evidence suggests that inclusion of test taker criticism in the question review process results in significantly higher educator self-efficacy after controlling for the educator self-efficacy pretest scores. Also, this study revealed strong similarities between the question review quality of the control group and the experimental groups. From each domain (math content, construct-irrelevance, cognitive psychology, and consequences of testing) analyzed, the group discussion and recommendation similarities far outnumbered the differences. In other words, inclusion of test taker criticism in the question review process does not result in an increase in question review quality based on the data collected in this study.

For this study I developed a self-efficacy instrument which can be used in future question reviews to evaluate how the question review training or participation impacted the educators' self-efficacy. This study also shed lights on the impact of test taker criticism on educators' test review process. That is, although test taker criticism can help increase educator self-efficacy, it did not essentially improve their question review quality. The implication would be that experienced educators can continue to review questions on behalf of test takers if the cost of including test taker criticism in the question review process is prohibitive.

# APPENDICES

## APPENDIX A

## Consent Forms

University of Illinois at Chicago
Research Information and Consent for Participation
The impact of test-taker criticism on educators'
self-efficacy and the detection and resolution of question issues
Student Form

You are being asked to participate in a research study. Researchers are required to provide a consent form such as this one to tell you about the research, to explain that taking part is voluntary, to describe the risks and benefits of participation, and to help you to make an informed decision. You should feel free to ask the researchers any questions you may have.

Principal Investigator Name and Title: Bill Trapp, PhD Candidate
Department and Institution: Educational Psychology Department, University of Illinois at Chicago
Address and Contact Information: 5668 IL Route 38, Dekalb, IL 60115
Email: wtrapp2@uic.edu
Phone: (815) 217-4143

Faculty Sponsor: Yue Yin, Ph.D., Associate Professor
Department and Institution: Educational Psychology Department, University of Illinois at Chicago
Address and Contact Information: 1040 W. Harrison St. MC 147 Chicago, IL 60607
Email: yueyin@uic.edu
Phone: (765) 430-3545

**Why am I being asked?**

You are being asked to be a participant in a research study about college admission test questions and the influence of test taker comments on educator reviewers and educator ability to identify and resolve issues. You have been asked to participate in the research because you are a current or former student at East Kentwood High School, you are 18- or 19-years old, and you have taken the SAT college admissions test recently.

Your participation in this research is voluntary. Your decision whether or not to participate will not affect your current or future dealings with East Kentwood High School or the University of Illinois at Chicago. If you decide to participate, you are free to withdraw at any time without affecting that relationship. You will receive no compensation if you withdraw, and you must complete all activities to receive the reimbursement described in this document. A total of eight students may be involved in this research.

**What is the purpose of this research?**

Each year, millions of high school students take college admissions tests. During the test development process, teachers evaluate test questions and make recommendations to improve the test questions. The purpose of this study is to investigate how student evaluation of test questions influences the recommendations of high school and college faculty in their review of the same questions.

**What procedures are involved?**
Participation in this study will involve the following procedures:

**Group Discussion.** On a Saturday morning, a total of 8 students will meet together and review college admission test questions and recommend changes to improve the quality of the questions. The discussion will last no more than 3 hours. The discussion will be audio recorded and all audio files will be destroyed within 48 hours of being transcribed.

**Survey.** After the completion of the group discussion, you will complete a written survey. Questions on the survey will explore attitudes and perceptions of the group discussion.
In summary, each participant will participate in a group discussion and complete one survey. All data will be collected on a single day.

**What are the potential risks and discomforts?**
The research has minimal risks to you. The East Kentwood High School administration is aware that I am requesting your participation in this research. There is a risk that other students from East Kentwood High School will know that you are a research subject. To the best of our knowledge, the things you will be doing have no more risk of harm than you would experience in everyday life. All data collected during the study will remain confidential; any identifying information will be deleted from any information disseminated and all data will be aggregated. No one, including the participants' school and family, will have access to the data other than the researcher. There is the risk that a breach of privacy and confidentiality may occur.

**Are there benefits to taking part in the research?**
There are no direct benefits to taking part in this research. The knowledge gained from this study may improve your understanding of assessment. Therefore, you may better understand your test scores and be able to explain how standardized tests are developed to others. The findings will extend research on test validity evidence and could assist test developers in improving test question review guidelines.

**What other options are there?**
Participation in this study is voluntary. You may decline to participate in this study.

**What about privacy and confidentiality?**
The people who will know that you are a research subject are the researchers, other participants. Otherwise information about you will only be disclosed to others with your written permission, or if necessary to protect your rights or welfare or if required by law.

When the results of the research are published or discussed in conferences, no information will be included that would reveal your identity. No one other than the researchers will have access to the data, including the participants' school or family. Each participant will be assigned a number. The list with the participants' contact information and linked number will be stored on the researcher's desktop in a password-protected file. All additional data collected will be de-identified and kept separate from the contact list. All de-identified electronic data will be stored on a password-protected desktop computer and all de-identified hard copy data will be stored in a locked file cabinet in the researcher's office. Once all data is collected, the contact list and codes will be destroyed. Additionally, audio files will be destroyed within 48 hours of being transcribed. All other data will be stored until the study is completed.

**What are the costs for participating in this research?**
There are no costs to you for participating in this research.

**Will I be reimbursed for any of my expenses or paid for my participation in this research?**
Upon completion of the survey, participants will receive a $25 gift card.

**Can I withdraw or be removed from the study?**
If you decide to participate, you are free to withdraw your consent and discontinue participation at any time without penalty. If you discontinue participating in any of the procedures involved, you will not be able to continue participating in the research study. The investigator may withdraw you from this research without your consent if circumstances arise which warrant doing so (e.g., not responding to verbal or written questions). You will receive no compensation if you withdraw, and you must complete all activities to receive the reimbursement described in this document.

**Who should I contact if I have questions?**
If you have any questions about this study or your part in it, or if you have questions, concerns or complaints about the research you may contact the primary researcher, Bill Trapp, at (815) 217-4143 or at wtrapp2@uic.edu. Additionally, you may contact the researcher's faculty sponsor, Dr. Yue Yin, at yueyin@uic.edu or(765) 430-3545.

**What are my rights as a research subject?**
If you feel you have not been treated according to the descriptions in this form, or if you have any questions about your rights as a research subject, including questions, concerns, complaints, or to offer input, you may call the Office for the Protection of Research Subjects (OPRS) at (312) 996-1711 or 1-866-789-6215 (toll-free) or e-mail OPRS at uicirb@uic.edu.

**Remember:**
Your participation in this research is voluntary. Your decision whether or not to participate will not affect your current or future relations with the East Kentwood High School or the University of Illinois. If you decide to participate, you are free to withdraw at any time without affecting that relationship. You will receive no compensation if you withdraw, and you must complete all activities to receive the reimbursement described in this document.

## APPENDIX A (continued)

**Signature of Subject or Legally Authorized Representative**

I have read (or someone has read to me) the above information. I have been given an opportunity to ask questions and my questions have been answered to my satisfaction. I agree to participate in this research. I will be given a copy of this signed and dated form.

| | |
|---|---|
| _____ | _____ |
| Signature | Date |
| _____ | |
| Printed Name | |
| _____ | _____ |
| Signature of Person Obtaining Consent | Date (must be same as subject's) |
| _____ | |
| Printed Name of Person Obtaining Consent | |

## APPENDIX A (continued)

University of Illinois at Chicago
Research Information and Consent for Participation
The impact of test-taker criticism on educators'
self-efficacy and the detection and resolution of question issues
Educator Form

You are being asked to participate in a research study. Researchers are required to provide a consent form such as this one to tell you about the research, to explain that taking part is voluntary, to describe the risks and benefits of participation, and to help you to make an informed decision. You should feel free to ask the researchers any questions you may have.

Principal Investigator Name and Title: Bill Trapp, PhD Candidate
Department and Institution: Educational Psychology Department, University of Illinois at Chicago
Address and Contact Information: 5668 IL Route 38, Dekalb, IL 60115
Email: wtrapp2@uic.edu
Phone: (815) 217-4143

Faculty Sponsor: Yue Yin, Ph.D., Associate Professor
Department and Institution: Educational Psychology Department, University of Illinois at Chicago
Address and Contact Information: 1040 W. Harrison St. MC 147 Chicago, IL 60607
Email: yueyin@uic.edu
Phone: (765) 430-3545

**Why am I being asked?**

You are being asked to be a participant in a research study about college admission test question and the influence of test taker comments on educator reviewers and resulting question quality. You have been asked to participate in the research because you are high school or college-level math educator in the Grand Rapids area. Your participation in this research is voluntary. Your decision whether or not to participate will not affect your current or future dealings with East Kentwood High School or the University of Illinois at Chicago. If you decide to participate, you are free to withdraw at any time without affecting that relationship. A total of sixteen math teachers may be involved in this research.

**What is the purpose of this research?**

Each year, millions of high school students take college admissions tests. During the test development process, teachers evaluate test questions and make recommendations to improve the test questions. The purpose of this study is to investigate how student evaluation of test questions influences the recommendations of high school and college faculty in their review of the same questions.

**What procedures are involved?**

You will participate in an experimental study with two conditions. Participants in condition 1 will review test questions in a focus group setting with other educators. Participants in condition 2 will review the same questions in the same setting, and also have access to student evaluative

comments about the test questions. Participation in this study will involve the following procedures:

**Group Discussion.** On a Saturday, you will review college admission math test questions with seven other educators and recommend changes to improve the quality of the questions. The discussion will last no longer than 4 hours. The discussion will be audio recorded and all audio files will be destroyed within 48 hours of being transcribed.

**Surveys.** Upon consent to participating, you will be asked to complete two surveys electronically. The purpose of the first survey is to collect demographic information about each participant. The second survey is a self-reporting measure of educator self-efficacy regarding your confidence to complete the group discussion task. Each survey will take approximately 15 minutes. At the completion of the group discussion, you will be asked to complete an additional two surveys electronically. The third survey is identical to the second survey. The purpose of the fourth survey is to collect your feedback on the process so that future data collection events may be improved.

In summary, each participant will participate in a group discussion and complete four surveys.

**What are the potential risks and discomforts?**
The research has minimal risks to you. There is a risk that educators in the Grand Rapids area will know that you are a research subject. To the best of our knowledge, the things you will be doing have no more risk of harm than you would experience in everyday life. All data collected during the study will remain confidential; any identifying information will be deleted from any information disseminated and all data will be aggregated. No one, including the participants' employer or supervisor, will have access to the data other than the researcher. There is the risk that a breach of privacy and confidentiality may occur.

**Are there benefits to taking part in the research?**
There are no direct benefits to taking part in this research. The knowledge gained from this study may improve your understanding of assessment. By participating in the study, you may be able to reflect on the way you evaluate test questions and increase both your self-efficacy in the task and the resulting quality of the test questions. The findings will extend research on test validity evidence and could assist test developers in improving test question review guidelines.

**What other options are there?**
Participation in this study is voluntary. You may decline to participate in this study.

**What about privacy and confidentiality?**
The people who will know that you are a research subject are the researchers, other participants. Otherwise information about you will only be disclosed to others with your written permission, or if necessary to protect your rights or welfare or if required by law.

When the results of the research are published or discussed in conferences, no information will be included that would reveal your identity. No one other than the researchers will have access to the

data, including the participants' school or family. Each participant will be assigned a number. The list with the participants' contact information and linked number will be stored on the researcher's desktop in a password-protected file. All additional data collected will be de-identified and kept separate from the contact list. All de-identified electronic data will be stored on a password-protected desktop computer and all de-identified hard copy data will be stored in a locked file cabinet in the researcher's office. Once all data is collected, the contact list and codes will be destroyed. Additionally, audio files will be destroyed within 48 hours of being transcribed. All other data will be stored until the study is completed.

**What are the costs for participating in this research?**
There are no costs to you for participating in this research.

**Will I be reimbursed for any of my expenses or paid for my participation in this research?**
Upon completion of the survey, participants will receive a $50 gift card.

**Can I withdraw or be removed from the study?**
If you decide to participate, you are free to withdraw your consent and discontinue participation at any time without penalty. If you discontinue participating in any of the procedures involved, you will not be able to continue participating in the research study. The investigator may withdraw you from this research without your consent if circumstances arise which warrant doing so (e.g., not responding to verbal or written questions).

**Who should I contact if I have questions?**
If you have any questions about this study or your part in it, or if you have questions, concerns or complaints about the research you may contact the primary researcher, Bill Trapp, at (815) 217-4143 or at wtrapp2@uic.edu. Additionally, you may contact the researcher's faculty sponsor, Dr. Yue Yin, at yueyin@uic.edu or (765) 430-3545.

**What are my rights as a research subject?**
If you feel you have not been treated according to the descriptions in this form, or if you have any questions about your rights as a research subject, including questions, concerns, complaints, or to offer input, you may call the Office for the Protection of Research Subjects (OPRS) at (312) 996-1711 or 1-866-789-6215 (toll-free) or e-mail OPRS at uicirb@uic.edu.

**Remember:**
Your participation in this research is voluntary. Your decision whether or not to participate will not affect your current or future relations with the East Kentwood High School or the University of Illinois. If you decide to participate, you are free to withdraw at any time without affecting that relationship.

**Signature of Subject or Legally Authorized Representative**
I have read (or someone has read to me) the above information. I have been given an opportunity to ask questions and my questions have been answered to my satisfaction. I agree to participate in this research. I will be given a copy of this signed and dated form.

**APPENDIX A (continued)**

| | |
|---|---|
| _____ | _____ |
| Signature | Date |
| _____ | |
| Printed Name | |
| _____ | _____ |
| Signature of Person Obtaining Consent | Date (must be same as subject's) |
| _____ | |
| Printed Name of Person Obtaining Consent | |

## APPENDIX B

### Teacher Background Survey

Q1 Participant ID Number: _____

Q2 Sex / Gender (optional): _____

Q3 Race/Ethnicity (optional). _____

Q4 Please list all educational degree(s) you have obtained and your major/field.

Q5 Please list all current K-12 teaching endorsements which you are certified to teach:

Q6 For each school/college/university you have taught at, list the **name of the school and the number of years you taught at that school**.

Q7 During the past three years, please list the names of the courses you have taught and describe the students in the course (high school freshman, college juniors, remedial college course)

Q8 Please describe any experience you have had reviewing standardized test questions prior to this event.

## APPENDIX C

## SAT Question Review

### Question 1

*Alignment:* HSF-TF.2. Explain how the unit circle in the coordinate plane enables the extension of trigonometric functions to all real numbers, interpreted as radian measures of angles traversed counterclockwise around the unit circle.
*Key: C*
*Test: SAT*
*Calculator usage: NO CALC*
*Rigor: Fluency*
*Difficulty: Easy*

If $\cos \theta = 0.5$, which of the following is also equal to 0.5?

A. $\sin \theta$
B. $\tan \theta$
C. $\sin (90 - \theta)$
D. $\tan (90 - \theta)$

## APPENDIX C (continued)

### Question 2

*Alignment:* HSG-GPE.1. Derive the equation of a circle of given center and radius using the Pythagorean Theorem; complete the square to find the center and radius of a circle given by an equation.
*Key: D*
*Test: SAT*
*Calculator usage: NO CALC*
*Rigor: Fluency*
*Difficulty: Medium*

In the *xy*-plane, what is the radius of the circle with equation $x^2 - 6x + y^2 + 4y = 8$?

A. $\sqrt{3}$
B. $2\sqrt{2}$
C. $\sqrt{13}$
D. $\sqrt{21}$

## APPENDIX C (continued)

### Question 3

*Alignment:* HSA-REI.4b. Solve quadratic equations by inspection (e.g., for $x^2 = 49$), taking square roots, completing the square, the quadratic formula and factoring, as appropriate to the initial form of the equation. Recognize when the quadratic formula gives complex solutions and write them as $a \pm bi$ for real numbers $a$ and $b$ (limited here to fluency in solving quadratic equations with diverse initial forms).
*Key: D*
*Test: SAT*
*Calculator usage: CALC*
*Rigor: Conceptual understanding*
*Difficulty: Medium*

If the equation $x^2 - 25x + c = 0$ has only one real solution, what is the value of c?

A. 5
B. $\frac{25}{2}$
C. $\frac{125}{4}$
D. $\frac{625}{4}$

## APPENDIX C (continued)

### Question 4

*Alignment:* HSF-IF.2. Use function notation, evaluate functions for inputs in their domains, and interpret statements that use function notation in terms of a context.
*Key: C*
*Test: SAT*
*Calculator usage: CALC*
*Rigor: Fluency*
*Difficulty: Hard*

If $f(x) = \frac{x(x-1)}{2}$, which of the following is true about $f(x-1)$?

A. $f(x - 1) = 1 - x + f(x)$
B. $f(x - 1) = f(x) - f(1)$
C. $f(x - 1) = (x - 1)f(x)$
D. $f(x - 1) = \frac{f(x) - f(1)}{2}$

## Question 5

*Alignment:* HSG-GMD.3. Use volume formulas for cylinders, pyramids, cones, and spheres to solve problems.
*Key: 51.3*
*Test: SAT*
*Calculator usage: CALC*
*Rigor: Application*
*Difficulty: Hard*

A gas tank manufacturer makes its tanks in the shape of a cylinder with hemispheres on both ends as shown.

The diameter of the cylinder and hemisphere is always 40 inches, but the length of the cylinder can vary. If the manufacturer wants to make a tank that has a volume of 100,000 cubic inches, how long should the cylinder part of the tank measure, to the closest tenth of an inch?

## APPENDIX C (continued)

### Question 6

*Alignment:* HSA-CED.1. Create equations and inequalities in one variable and use them to solve problems. Include equations arising from linear and quadratic functions, and simple rational and exponential function.
*Key: B*
*Test: SAT*
*Calculator usage: NO CALC*
*Rigor: Application*
*Difficulty*: Easy

Aaron is staying at a hotel that charges $49.95 per night. The price does not include a tax of 8% on the price of the room and a one-time reservation fee of $1.00. Which of the following represents the price Aaron will pay if $x$ is the number of nights he will spend at the hotel?

A. $(49.95 + 0.08x) + 1$
B. $1.08(49.95x) + 1$
C. $1.08(49.95x + 1)$
D. $1.08(49.95 + 1)x$

## APPENDIX C (continued)

### Question 7

*Alignment:* HSA-CED.4. Rearrange formulas to highlight a quantity of interest, using the same reasoning as in solving equations. For example, rearrange Ohm's law V = IR to highlight resistance R.
*Key: C*
*Test: SAT*
*Calculator usage: NO CALC*
*Rigor: Conceptual understanding*
*Difficulty: Medium*

A driver presses on the gas pedal of a car and the car accelerates. After an initial surge, the car accelerates at a constant rate from 4 to 10 seconds. The acceleration, $a$, of the car during this time can be found with the formula $a = (\frac{v_1 + v_2}{2})t$, where $v_1$ and $v_2$ are the initial and final speeds, respectively, and $t$ is the amount of time that elapsed between $v_1$ and $v_2$. Which equation represents the time elapsed in terms of the other variables?

A. $t = 2a(v_1 + v_2)$

B. $t = a\left(\frac{v_1 + v_2}{2}\right)$

C. $t = \frac{2a}{v_1 + v_2}$

D. $t = \frac{v_1 + v_2}{2a}$

## APPENDIX C (continued)

### Question 8

*Alignment:* HSA-REI.3. Solve linear equations and inequalities in one variable, including equations with coefficients represented by letters.
*Key: Any value between -3/4 and 3*
*Test: SAT*
*Calculator usage: NO CALC*
*Rigor: Fluency*
*Difficulty: Medium*

If $-1 < -3t + 1 < \frac{1}{4},$ what is one possible value of $9t - 3$?

## APPENDIX C (continued)

### Question 9

*Alignment:* HSA-REI.6. Solve systems of linear equations exactly and approximately (e.g., with graphs), focusing on pairs of linear equations in two variables.
*Key: 9/13 or .692*
*Test: SAT*
*Calculator usage: NO CALC*
*Rigor: Fluency*
*Difficulty: Hard*

Given the system of equations:

$2x + 3y = 27$

$y = -5x + 12$

What is the $x$-value of solution $(x, y)$ to the system of equations above?

## APPENDIX C (continued)

### Question 10

*Alignment:* HSF-BF.1. Write a function that describes a relationship between two quantities.
*Key: A*
*Test: SAT*
*Calculator usage: NO CALC*
*Rigor: Application*
*Difficulty: Medium*

Gene uses the following table to estimate the profit made based on the number of people that dine in his restaurant each day.

| Number of People | Estimated Profit |
|---|---|
| 15 | ($285) loss |
| 30 | $145 |
| 45 | $575 |
| 60 | $1005 |
| 75 | $1435 |

Which linear function represents the estimated profit, $p(x)$, based on the number of people, $x$, who dined in the restaurant?

A. $p(x) = \frac{86}{3}x - 715$

B. $p(x) = \frac{86}{3}x - 285$

C. $p(x) = 430x - 715$

D. $p(x) = 430x - 285$

## APPENDIX C (continued)

### Question 11

*Alignment:* HSS-ID.7. Interpret the slope (rate of change) and the intercept (constant term) of a linear model in the context of the data.
*Key: C*
*Test: SAT*
*Calculator usage: NO CALC*
*Rigor: Application*
*Difficulty: Medium*

Jan is a lawyer who is researching legal cases related to water pollution in her area. Her initial search includes cases submitted from 1965 through 2000. The equation $f(x) = 11.5x + 14$ can be used to estimate the number of cases, $f(x)$, where $x$ is the number of years since 1964. What is the meaning of the constant 11.5 in this equation?

A. The number of cases in 1965.
B. The number of cases in 2000.
C. The increase in cases per year.
D. The increase in cases from 1965 to 2000.

## APPENDIX C (continued)

### Question 12

*Alignment:* HSF-LE.2. Construct linear and exponential functions, including arithmetic and geometric sequences, given a graph, a description of a relationship, or two input-output pairs (include reading these from a table).
*Correct answer: 1/17 or .059*
*Test: SAT*
*Calculator usage: CALC*
*Rigor: Application*
*Difficulty: Hard*

According to a census, the population of a town doubled approximately every 17 years since 1950. If the equation $P = P_0 2^{kt}$, where $P_0$ is the population of the town in 1950, will be used to model the population of the town $t$ years after 1950, what should the value of $k$ be?

## Question 13

*Alignment:* HSA-SSE.1. Interpret expressions that represent a quantity in terms of its context. a) Interpret parts of an expression, such as terms, factors, and coefficients. b) Interpret complicated expressions by viewing one or more of their parts as a single entity. For example, interpret $P(1 + r)^n$ as the product of P and a factor not depending on P.
*Key: B*
*Test: SAT*
*Calculator usage: NO CALC*
*Rigor: Application*
*Difficulty: Hard*

$$\frac{1}{x} + \frac{2}{x} = \frac{1}{5}$$

Anise needs to complete a printing job and she will be using both printers in her office for it. One of the printers is twice as fast as the other and together they can finish the job in 5 hours. Anise wants to figure out how long would have taken her to finish the job if she had used the slower printer only. For this, she writes the equation above. What does the expression $\frac{1}{x}$ in this equation represent?

A. The time it takes the slower printer to finish the printing job.
B. The portion of job that the slower printer would complete in one hour.
C. The portion of job that the faster printer would complete in two hours.
D. The time it takes the slower printer to complete 1/5 of the printing job.

## Question 14

*Alignment:* HSA-SSE.2. Use the structure of an expression to identify ways to rewrite it. For example, see $x^4 – y^4$ as $(x^2)^2 – (y^2)^2$, thus recognizing it as a difference of squares that can be factored as $(x^2 – y^2)(x^2 + y^2)$.
*Key: D*
*Test: SAT*
*Calculator usage: NO CALC*
*Rigor: Fluency*
*Difficulty: Hard*

The expression $\frac{1}{\sqrt{2.5} – \sqrt{0.4}}$ is equal to which of the following?

A. $\frac{1}{0.3}$

B. $\frac{1}{\sqrt{3}}$

C. $\frac{3}{\sqrt{10}}$

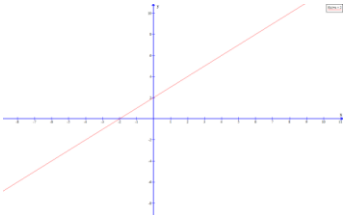
D. $\frac{\sqrt{10}}{3}$

## Question 15

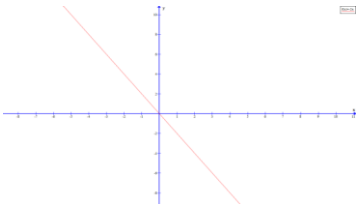
*Alignment:* HSA-REI.10. Understand that the graph of an equation in two variables is the set of all its solutions plotted in the coordinate plane, often forming a curve (which could be a line).
*Key: B*
*Test: SAT*
*Calculator usage: CALC*
*Rigor: Conceptual understanding*
*Difficulty: Medium*

Which of the following could be the graph of $y = mx$ in the $xy$-plane, where m is negative?

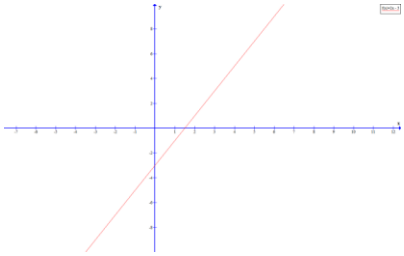
A.



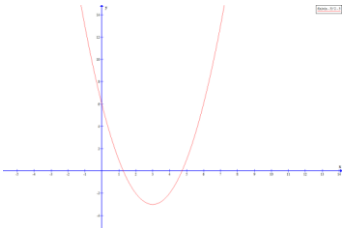
B.



C.



D.



## Question 16

*Alignment:* HSA-APR 3. Identify zeros of polynomials when suitable factorizations are available, and use the zeros to construct a rough graph of the function defined by the polynomial.
*Key: A*
*Test: SAT*
*Calculator usage: CALC*
*Rigor: Application*
*Difficulty: Medium*

$$y = x^3 + 2x^2 + x$$

If the equation shown is graphed on the *xy*-plane, which of the following points will the curve pass through?

A. (0,0)
B. (1,0)
C. (2,0)
D. (3,0)

## APPENDIX C (continued)

The table below displays voting information gathered by the US Census Bureau after the November 2012 presidential election.

**Reported Voting by Age (in thousands)**

| | | Voted | Did Not Vote | No Response | Total |
|---|---|---|---|---|---|
| **Age** | **18 to 34 Years** | 30,329 | 23,211 | 9,468 | 63,008 |
| | **35 to 54 Years** | 47,085 | 17,721 | 9,476 | 74,282 |
| | **55 to 74 Years** | 43,075 | 10,092 | 6,831 | 59,998 |
| | **75 Years and Over** | 12,459 | 3,508 | 1,827 | 17,794 |
| | **Total** | 132,948 | 54,532 | 27,602 | 215,082 |

### Question 17

*Alignment:* HSS-ID.5. Summarize categorical data for two categories in two-way frequency tables. Interpret relative frequencies in the context of the data (including joint, marginal, and conditional relative frequencies). Recognize possible associations and trends in the data.
*Key: C*
*Test: SAT*
*Calculator usage: CALC*
*Rigor: Application*
*Difficulty: Easy*

Based on the data in the table, members of which age range are most likely to report that they voted?
A. 18 to 34 year olds
B. 35 to 54 year olds
C. 55 to 74 year olds
D. People over 75 years old

### Question 18

*Alignment:* HSS-IC.1. Understand statistics as a process for making inferences about population parameters based on a random sample from that population.
*Key: C*
*Test: SAT*
*Calculator usage: CALC*
*Rigor: Application*
*Difficulty: Medium*

If the population of US citizens of voting age grows by 30 million, using the data in the table, what would be the approximate number of voters?
A. 151,460,000
B. 162,948,000
C. 217,480,000
D. 245,082,000

## APPENDIX C (continued)

### Questions 19 and 20

*Alignments:*

- 6.RP.A.3b Solve unit-rate problems including those involving unit pricing and constant speed. For example, if it took 7 hours to mow 4 lawns, then at that rate, how many lawns could be mowed in 35 hours? At what rate were lawns being mowed?
- 6.RP.A.3c Find a percent of a quantity as a rate per 100 (e.g., 30% of a quantity means 30/100 times the quantity); solve problems involving finding the whole, given a part and the percent.
- 6.RP.A.3d Use ratio reasoning to convert measurement units; manipulate and transform units appropriately when multiplying or dividing quantities.
- 7.RP.A.1 Compute unit rates associated with ratios of fractions, including ratios of lengths, areas and other quantities measured in like or different units. For example, if a person walks 1/2 mile in each 1/4 hour, compute the unit rate as the complex fraction 1/2/1/4 miles per hour, equivalently 2 miles per hour.
- 7.RP.A.2b Identify the constant of proportionality (unit rate) in tables, graphs, equations, diagrams, and verbal descriptions of proportional relationships.
- 7.RP.A.3 Use proportional relationships to solve multistep ratio and percent problems. Examples: simple interest, tax, markups and markdowns, gratuities and commissions, fees, percent increase and decrease, percent error.

*Key: #19: 14070; #20: 73667*
*Test: SAT*
*Calculator usage: CALC*
*Rigor: Application*
*Difficulty: #19: Easy; #20: Hard*

Larissa is planning to start a business. She has outlined the initial costs involved in the table below.

| **Expense** | **Cost** |
|---|---|
| Registration of company | $129 (annual fee) |
| Rent and utilities | $1300 (monthly fee) |
| Salaries | $3000 (monthly fee) |
| Employer taxes | 13.3% of salaries paid quarterly |

Larissa decides that she needs at least 9 months of funding before her company will make enough money to operate on its own. She ends up taking out a loan that covers one year of her expenses, the interest on the loan is 0.5% compounded monthly.

### Question 19

How much more money did Larissa get than what she needed?

### Question 20

Larissa wants to pay the entire loan back in one payment at the end of 3 years and expects to double the amount she pays in salaries at the end of 2 years. What will be the average yearly revenue her company needs to make in this time in order for her company to continue operating and these conditions to be met? Give your answer to the nearest dollar.

## APPENDIX D

## Rating Form

---

**Individual Review Comments**

Q1 Participant ID Number: _____

Q2 Test Question Number: _____

Q3 Initial Recommendation (check one):

- ○ Accept as is
- ○ Accept with revisions
- ○ Reject

Q4 Initial Comments:

| |
|---|
| |

---

**Comments after Group Discussion**

Q5 Recommendation after group discussion (check one):

- ○ Accept as is
- ○ Accept with revisions
- ○ Reject

Q6 Do you agree with the group recommendation? Please explain.

| |
|---|
| |

---

## APPENDIX E

## Appraisal Inventory

In the process of developing the SAT test, the test development committee participants evaluate each question using various criteria. This questionnaire lists some of those criteria and considerations. The questionnaire is designed to help us gain a better understanding of how confident reviewers feel evaluating a question based on each criterion. Please rate how certain you are that you can do the things below by selecting the appropriate rating. Your answers will be kept strictly confidential and will not be identified by name.

Q1. Participant ID Number:_____

Q2 **Rate your degree of confidence in evaluating a question's content based on each of the following statements (place an "X" in one column, for each row).**

| Statement | Highly unsure | Unsure | Neutral | Confident | Highly confident |
|---|---|---|---|---|---|
| The question's clarity, including ambiguity or imprecision in a question | | | | | |
| The question's mathematical precision | | | | | |
| The question's defensibility from public criticism | | | | | |
| The question's difficulty, what percent of test takers will answer a question correctly | | | | | |
| The question's discrimination, whether high ability test takers answer a question correctly more often than low ability test takers | | | | | |
| The question's alignment to my state standards | | | | | |
| The question's cognitive alignment (e.g. Bloom's taxonomy, Webb Depth of Knowledge) | | | | | |
| The question's alignment to the SAT content descriptions | | | | | |
| The question's relevance to the to the purposes of the SAT | | | | | |
| The question's adherence to SAT style and formatting rules | | | | | |
| The question's conformity with the SAT test directions | | | | | |

## APPENDIX E (continued)

Q3 Questions may be asked in a way that is unfair to one or more subgroups of test takers.
**Rate your degree of confidence in evaluating a question's unfairness to test takers as a result of each of the following statements (place an "X" in one column, for each row).**

| Statement | Highly unsure | Unsure | Neutral | Confident | Highly confident |
|---|---|---|---|---|---|
| There is unfairness due to unrelated knowledge or skills required to answer the question | | | | | |
| There is unfairness due to a context which evokes one or more distracting emotions | | | | | |
| There is unfairness due to physical abilities described and/or the physical question format | | | | | |
| There is unfairness due to a context impacting test taker motivation | | | | | |
| There is unfairness due to a context impacting test anxiety | | | | | |
| There is unfairness due to familiarity with question format | | | | | |
| There is unfairness due to text complexity (e.g. vocabulary, sentence structure, sentence length, etc…) | | | | | |
| There is unfairness due to the test taker's age | | | | | |
| There is unfairness due to the test taker's sex and/or gender identity | | | | | |
| There is unfairness due to the test taker's race and/or ethnicity | | | | | |
| There is unfairness due to the test taker's home (geographic region within the US) | | | | | |
| There is unfairness due to the test taker's home (urban or rural) | | | | | |
| There is unfairness due to the test taker's home (US or international) | | | | | |
| There is unfairness due to the test taker's first and/or primary language | | | | | |
| There is unfairness due to the test taker's socioeconomic status | | | | | |
| There is unfairness due to the test taker's religion | | | | | |

## APPENDIX E (continued)

Q4 Based on the content and presentation of a question, educators can hypothesize what mental processes test takers are likely to utilize.

**Rate your degree of confidence in evaluating the appropriateness of a question based on each of the following mental processes test takers are likely to utilize.**

| Statement | Highly unsure | Unsure | Neutral | Confident | Highly confident |
|---|---|---|---|---|---|
| the question's response format (MC or SPR) provides insight into the student's thought process(es) | | | | | |
| the student's thought process(es) to answer the question correctly are worth the additional raw score point they will earn | | | | | |
| marking any of the incorrect answers is worth no additional raw score points | | | | | |
| an appropriate amount of working (or short-term) memory was required in order to answer the question | | | | | |
| an appropriate amount of long-term memory was required in order to answer the question | | | | | |
| a question differentiates between test takers who use weak problem-solving methods (e.g. trial and error) and test takers who use strong methods (e.g. standard algorithms, mathematical reasoning) | | | | | |
| a question differentiates between test takers who can apply prior experience to a previously unsolved question and students who struggle to recognize opportunities to use prior experience in new situations | | | | | |
| a question creates an opportunity to differentiate between test takers who are content 'experts' and test takers who are content 'novices' | | | | | |
| a question creates an opportunity to differentiate between test takers who utilize metacognition (thinking about their thinking) while problem solving and those who do not | | | | | |

## APPENDIX E (continued)

Q5 Some questions may generate intended or unintended consequences.
**Rate your degree of confidence in evaluating a test question's appropriateness based on each of the following intended or unintended consequences.**

| Statement | Highly unsure | Unsure | Neutral | Confident | Highly confident |
|---|---|---|---|---|---|
| The question may cause test anxiety in test takers prior to taking the test. | | | | | |
| The question may change test taker motivation prior to, during, or after taking the test. | | | | | |
| The question may change test preparation efforts of test takers. | | | | | |
| The question may change the instruction delivered by educators who become familiar with the test. | | | | | |
| The question may change how test stakeholders (students, teachers, counselors, administrations) use the test scores. | | | | | |

## APPENDIX F

### Student and Educator Feedback Questionnaire

Q1 Participant ID Number (optional): _____

We are interested your feedback on the **training provided at the beginning of the focus group**.

Q2 What aspects of the training did you find most helpful?

Q3 Did you learn anything that you will be able to apply beyond this today's research study? Please explain.

Q4 What aspects of the training can we improve for future studies?

We are interested in your feedback on the **group discussion**.

Q5 Do you feel that we heard all of your concerns about the test questions? If not, what was missed?

Q6 What aspect(s) of the process can we improve so that we obtain the highest quality test questions possible?

Q7 To what degree do you agree with this statement: "The recommendations from this group, when implemented, result in high quality test questions for the SAT."
Please explain.

Q8 Please share any other comments you have regarding the process used in this question review.

## APPENDIX G

### Summary of Test Taker Criticism

The comments in italic on the right are the consensus recommendations from test takers.

### Question 1

*Alignment:* HSF-TF.2. Explain how the unit circle in the coordinate plane enables the extension of trigonometric functions to all real numbers, interpreted as radian measures of angles traversed counterclockwise around the unit circle.
*Key: C*
*Test: SAT*
*Calculator usage: NO CALC*
*Rigor: Fluency*
*Difficulty: ~~Easy~~ Medium*

If cos $\theta = 0.5$, which of the following is also equal to 0.5?

A. sin $\theta$
B. tan $\theta$
C. sin $(90 - \theta)$
D. tan $(90 - \theta)$

*Accept with Revisions*

"I really don't remember this. But, when I was a junior and I took the SAT, I was in precalculus and this would have been fresh in my mind. So I don't think I would have struggled as much then."

"I thought that the difficulty as easy, but for advanced students it would be easy but most students it wouldn't be"

"I was in Algebra 2 last year, and yes we went over the unit circle but not, it wasn't in depth"

"I felt like it was easily comprehendible if you knew the content knowledge."

Very few students remembered how to solve this question. The discussion centered around the difficulty of the question. The students recommended to change the difficulty to "Medium" and they agreed to "Accept with Revisions" to the difficulty level.

On the rating form, 2 of the students disagreed with the recommendation, stating that the original difficulty level was appropriate. The remainder of the students agreed with the recommendation.

## APPENDIX G (continued)

### Question 2

*Alignment:* HSG-GPE.1. Derive the equation of a circle of given center and radius using the Pythagorean Theorem; complete the square to find the center and radius of a circle given by an equation.
*Key: D*
*Test: SAT*
*Calculator usage: NO CALC*
*Rigor: Fluency*
*Difficulty: Medium*

~~In the xy-plane, what is the radius of the circle with equation $x^2 - 6x + y^2 + 4y = 8$?~~
The standard equation of a circle is $(x–h)^2 + (y–k)^2 = r^2$. A circle has the equation $x^2 – 6x + y^2 + 4y = 8$. What is the radius of this circle?

A. $\sqrt{3}$
B. $2\sqrt{2}$ *Accept with Revisions*
C. $\sqrt{13}$
D. $\sqrt{21}$

"I think some students would miss the fact that they need to know the Pythagorean theorem. So maybe the question. Not saying that the question should have a hint, like say use it. But some word using the Pythagorean theorem because some kids are going to going to be like I don't even know what to do."
"I remember learning it. But then after that, teachers and peers would just show us ways to do it in our calculators so we didn't actually write it out like that. So I think putting it on the calculator… I don't think that's right. Then at the same time I guess it's not the same for everybody at other schools. I know here my teacher taught me ways to do it on my calculator."
"I think if you added to the calculator portion though, the test, there would be other ways to solve it"
"Like, geometry was 8th grade. Um, I took it then and really haven't done the specific things since, so it's pretty useless in that sense."
"I feel like you could keep this if you just give a hint on how to solve it. We just don't go over stuff like this enough in high school."

None of the students knew how to solve this question. We discussed when they remembered learning how to complete the square and they hypothesized why they forgot this skill. I demonstrated how to solve the question and after prompting them with the initial steps, many remembered the subsequent steps. The students requested a reminder and the revision shown above was presented.

The students approved the change and thought that the revised question would have a Medium difficulty. They agreed to "Accept with Revisions" to the question text. On the rating form, all students indicated agreement with the recommendation.

## APPENDIX G (continued)

### Question 3

*Alignment:* HSA-REI.4b. Solve quadratic equations by inspection (e.g., for $x^2 = 49$), taking square roots, completing the square, the quadratic formula and factoring, as appropriate to the initial form of the equation. Recognize when the quadratic formula gives complex solutions and write them as $a \pm bi$ for real numbers $a$ and $b$ (limited here to fluency in solving quadratic equations with diverse initial forms).
*Key: D*
*Test: SAT*
*Calculator usage: CALC*
*Rigor: Conceptual understanding*
*Difficulty: Medium*

If the equation $x^2 - 25x + c = 0$ has only one real solution, what is the value of c?

A. 5
B. $\frac{25}{2}$
C. $\frac{125}{4}$ *Accept As Is*
D. $\frac{625}{4}$

There was a very brief discussion about this question. One student said the wording was clear another student made the following statement about the match to the alignment, "With the objective, or the alignment, they would pick one of those things to do."
I asked if students had seen a question like this before and they indicated that they had.
I asked if students had seen a question in this particular format, with c, and you needed to determine the value of c, and they indicated that they were familiar with the question format.
They agreed to "Accept As Is". On the rating form, all students indicated agreement with the recommendation.

## APPENDIX G (continued)

### Question 4

*Alignment:* HSF-IF.2. Use function notation, evaluate functions for inputs in their domains, and interpret statements that use function notation in terms of a context.
*Key: ~~C~~ A*
*Test: SAT*
*Calculator usage: CALC*
*Rigor: Fluency*
*Difficulty: Hard*

If $f(x) = \frac{x(x-1)}{2}$, which of the following is true about $f(x-1)$?

| | |
|---|---|
| ~~A. $f(x-1) = 1 - x + f(x)$~~ | $f(x-1) = ((x-1)(x-2))/2$ |
| ~~B. $f(x-1) = f(x) - f(1)$~~ | $f(x-1) = ((x-1)(x-1))/2$ |
| ~~C. $f(x-1) = (x-1)f(x)$~~ | $f(x-1) = (x-2)/2$ |
| ~~D. $f(x-1) = \frac{f(x) - f(1)}{2}$~~ | $f(x-1) = (x-1)/2$ |

*Accept with Revisions*

"I thought it would be ok on the no calculator section. I'm just saying a student might not use it, might not use a calculator."
"Not everyone is going to know how to that one so I thought that was a good fit and I thought that the difficulty was hard. All in all, I thought the question with the alignment was hard."
"I just would not know how to plug this in my calculator"
"I would end up having to solve it by hand."
"I saw that the answer was C, but I didn't know how to get there. And so like if I was taking the SAT and I saw this question I would just, honestly I would just have given up and either guessed or waited until the end and like whatever time I had left I would have tried to solve it."
"It's really confusing. Like, already having to manipulate it, I don't know"
"I mean the answers that are there make sense to me but I wouldn't know how to get there."
"I don't know about everyone else but I don't want to say that it should be rejected because I feel like it's over my head. But I don't want to reject it just because it is hard."
"I never had to go back and put it in that notation. That doesn't mean that's how other teachers work. And how other teachers might tell their kids put it back in this notation. So I guess it's just that I was never taught how to do it."

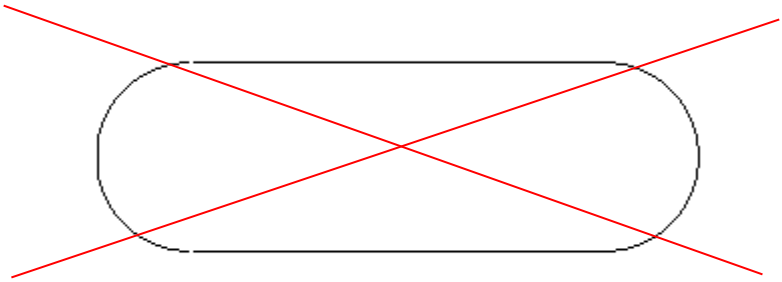
None of the students knew how to solve this question. They initially described the question as hard, but when prompted they confirmed that they did not know how to solve. They were hesitant to reject the question because they thought it matched the standard and because they thought students in other schools might be better prepared than they were. I asked them, if one of their teachers was participating on a committee like this and representing them, what would they hope the teacher would recommend? Several students hoped that the teacher would recommend rejecting the question.
I described how to solve the problem, and the suggestion was made to use answer choices that were obtained at an earlier step in the process. The proposed answer choices are shown above:
They agreed to "Accept with Revisions" to the changes to the answer choices and the key. On the rating form, all students indicated agreement with the recommendation, and one student indicated they would have been ok with the question without any changes.

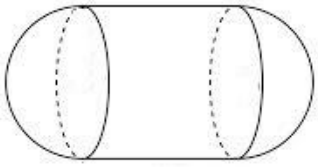
**APPENDIX G (continued)**

**Question 5**

*Alignment:* HSG-GMD.3. Use volume formulas for cylinders, pyramids, cones, and spheres to solve problems.
*Key: 51.3*
*Test: SAT*
*Calculator usage: CALC* *Accept with Revisions*
*Rigor: Application*
*Difficulty: Hard*

A gas tank manufacturer makes its tanks in the shape of a cylinder with hemispheres on both ends as shown.





The diameter of the cylinder and hemisphere is always 40 inches, but the length of the cylinder can vary. If the manufacturer wants to make a tank that has a volume of 100,000 cubic inches, how long should the cylinder part of the tank measure, to the closest tenth of an inch?

**"**When I took the SAT it comes with a formula sheet right?"
**"**Are the formulas for cylinders and columns and stuff on the formula sheet?"
**"**I feel like most people are gonna disregard that first sentence and like if there was a better visual. Like if it was 3D, then they'd be more likely to use those equations."
**"**The visual doesn't make sense. There aren't any. It looks two dimensional."
**"**Yeah, and the way it's worded with that picture. It's hard to see the cylinder and the hemisphere."
**"**With the visual there should probably be labels with the length and the diameter so it's easier for us to locate them."

The students first asked about the formula sheet and I confirmed that the formulas for the sphere and the right circular cylinder are provided. Several comments were made about the poor quality of the figure. We discussed, at length, ways to present the figure. We discussed what should or should not be labeled, and whether labeling certain aspects of the figure compromised the question alignment. The image shown at right was approved.
They agreed to "Accept with Revisions" to the figure. On the rating form, all students indicated agreement with the recommendation.

## APPENDIX G (continued)

### Question 6

*Alignment:* HSA-CED.1. Create equations and inequalities in one variable and use them to solve problems. Include equations arising from linear and quadratic functions, and simple rational and exponential function.
*Key: B*
*Test: SAT*
*Calculator usage: NO CALC*
*Rigor: Application*
*Difficulty*: Easy

Aaron is staying at a hotel that charges $49.95 per night. The price does not include a tax of 8% on the price of the room and a one-time reservation fee of $1.00. Which of the following represents the price Aaron will pay if $x$ is the number of nights he will spend at the hotel?
A. $(49.95 + 0.08x) + 1$
B. $1.08(49.95x) + 1$
C. $1.08(49.95x + 1)$
D. $1.08(49.95 + 1)x$

*Accept As Is*

"I thought it was easy to do. I thought a lot of students would be able to do it."
"I think that um, you don't create the equation for it, you're not really solving for anything. You're just really creating the equation and I don't know if that matters. So I saying that you should add a component to the question where you have to solve for the price if he stayed for a week. Like, I didn't even read the alignment when I read the question I just kind of did it. And I didn't think that I was doing the alignment"
"Yeah, I agree with that because maybe like at the end saying how much did you pay for some number of nights because then you'd be using that to solve the problem."
"I think that if you add it in, like how much did it cost for one night, then it might actually cause some kids not to create any equation. So I think, I mean I guess you are calculating and solving an equation if you solving it in a certain way in your head."
"I think it's fine as is. I was looking at it from the alignment and how students solve the problem. But I think we should definitely have questions where we make an equation and we should definitely have question that we need to find a specific answer."
"Yeah, I guess. I agree with that you shouldn't have questions about solving …but you could have questions that you were solving like how many nights, how much would it cost for this many nights? Like the fill in the blank ones where you do your numbers that you just plug in the value. Yeah, I think that's fine."

The students discussed the alignment at length. We discussed whether you must assess the entire alignment or if it is ok to assess just a portion of the alignment. They were able to solve this question, and they didn't have any other issues other than how to interpret the alignment.
They agreed to "Accept As is". On the rating form, all students indicated agreement with the recommendation and one student would have preferred that the question ask for a solution to a problem in addition to asking for the expression that represents the price Aaron will pay.

## APPENDIX G (continued)

### Question 7

*Alignment:* HSA-CED.4. Rearrange formulas to highlight a quantity of interest, using the same reasoning as in solving equations. For example, rearrange Ohm's law V = IR to highlight resistance R.
*Key: C*
*Test: SAT*
*Calculator usage: NO CALC*
*Rigor: Conceptual understanding*
*Difficulty: Medium*

~~A driver presses on the gas pedal of a car and the car accelerates. After an initial surge, the car accelerates at a constant rate from 4 to 10 seconds.~~ The acceleration, $a$, of ~~the~~ a car ~~during this time~~ can be found with the formula $a = (\frac{v_1 + v_2}{2})t$, where $v_1$ and $v_2$ are the initial and final speeds, respectively, and $t$ is the amount of time that elapsed between $v_1$ and $v_2$. Which equation represents the time elapsed in terms of the other variables?

A. $t = 2a(v_1 + v_2)$

B. $t = a\left(\frac{v_1 + v_2}{2}\right)$

*Accept with Revisions*

C. $t = \frac{2a}{v_1 + v_2}$

D. $t = \frac{v_1 + v_2}{2a}$

"Um, I thought it was pretty wordy for an SAT question. I think like maybe the first couple sentences aren't necessary. They're just kind of useless background information that doesn't really apply to what you…. So you could pretty much just delete that."

"Especially with the four to ten part. People may try to put that in the equation and it'd be distracting."

"Yeah, I think the first sentences are distracting and that's irrelevant"

The students knew how to solve this question. The discussion centered around how the first two sentences were not adding value. The students recommended to remove the first two sentences and make a modification to the third sentence. The revised question is shown above. They agreed to "Accept with Revisions". On the rating form, all students indicated agreement with the recommendation.

**APPENDIX G (continued)**

**Question 8**

*Alignment:* HSA-REI.3. Solve linear equations and inequalities in one variable, including equations with coefficients represented by letters.
*Key: Any value between ~~-3/4 and 3~~ 2/3 and 9/4*
*Test: SAT*
*Calculator usage: NO CALC*
*Rigor: Fluency*
*Difficulty: Medium*

If $-1 < -3t + 1 < \frac{1}{4}$, what is one possible value of ~~$9t - 3$~~ $t$?

*Accept with Revisions*

"I was confused"

"Yeah, me too"

"I wasn't sure like, how to even like, tackle the question."

"I would have just skipped it and gone on to a different question."

"Ok so I think I know this problem, what is a possible value of nine t minus three, but you're given negative three t plus one. Wouldn't you just? Nevermind nevermind."

"I feel like in general the problem is straight forward, if teachers go over this format."

"It's hard to correlate the two. Like nine t minus three"

None of the students knew how to solve this question. I demonstrated how to solve the question and after prompting them with the initial steps, many remembered the subsequent steps. They were hesitant to reject the question because they thought it matched the standard and because they thought students in other schools might be better prepared than they were. I proposed a simpler version of the problem shown above.
They agreed to "Accept with Revisions" to the question and the key. On the rating form, all students indicated agreement with the recommendation.

## APPENDIX G (continued)

### Question 9

*Alignment:* HSA-REI.6. Solve systems of linear equations exactly and approximately (e.g., with graphs), focusing on pairs of linear equations in two variables.
*Key: 9/13 or .692*
*Test: SAT*
*Calculator usage: NO CALC*
*Rigor: Fluency*
*Difficulty: Hard*

*Accept As Is*

Given the system of equations:

$2x + 3y = 27$

$y = -5x + 12$

What is the $x$-value of solution $(x, y)$ to the system of equations above?

"I thought this was a good question."

"I thought the difficulty should be medium. Cause I just remember this is the main thing I took out of algebra 2 and pre-calculus."

"Oh it's because it's already in the format y equals it would be easier to just plug it into the first equation. So if you wanted to keep it hard you could rearrange that second equation and then they'd have to go back into y equals and solve"

"When I was solving it, I just felt like it's one of those questions you get on the SAT. It's kind of like a break from one of the hard ones."

"Especially because it's on the no calculator portion, I'm sure there are other questions that are a lot harder on that section."

There was a very brief discussion about this question. The discussion focused on the difficulty, but ultimately the students did not feel the difficulty needed to change. Some students briefly hypothesized about the variety of SAT question difficulty and how this question fit into that distribution.
They agreed to "Accept As is". On the rating form, all students indicated agreement with the recommendation.

## APPENDIX G (continued)

### Question 10

Gene uses the following table to estimate the profit made based on the number of people that dine in his restaurant each day.

| Number of People | Estimated Profit |
|---|---|
| 15 | ~~($285) loss~~ –$285 |
| 30 | $145 |
| 45 | $575 |
| 60 | $1005 |
| 75 | $1435 |

*Accept with Revisions*

A linear function to represent this situation can be written in the form $p(x) = mx + b$, where m and b are constants. Which linear function represents the estimated profit, $p(x)$, based on the number of people, $x$, who dined in the restaurant?

A. $p(x) = \frac{86}{3}x - 715$

B. $p(x) = \frac{86}{3}x - 285$

C. $p(x) = 430x - 715$

D. $p(x) = 430x - 285$

"There were some troubles with this one."
"Yeah, I have no idea why that's the answer, at all."
"Yeah, I got D and I don't really know why"
"Yeah, the thing with this is that it's on a no calculator and the values are pretty big, I mean, in this situation would that, especially they have fractions here as well. I feel like it should be on the calculator portion. And the diagram, the negative profit, or I guess the loss, is in parentheses and with loss next to it and I feel it should just be negative. I mean I guess you could also keep loss there but the parentheses thing is kind of weird."
"I wasn't really sure about the question, like how it was stated. It was a little weird but. That's why I was confused. I was trying to figure out what to do."
"Yeah, I don't really think there's a problem with the question."
"It's just basic content knowledge"
"Yeah, I think that I just like overthought that you had to use a linear equation because it was based on the number of people. So I was trying to figure out. I don't know. I think it might be easier if you just asked write an equation based on the problem, the number of people who dined at the restaurant."
"If I would have sat longer and thought about it longer I think I would have gotten it."
"Maybe if you put a graph next to the table. Not even with data on it, to emphasize the fact that this is an equation, maybe? I don't know if that."

Several students struggled to solve this question. The conversation centered on finding a starting point, possibly because the question was too dense. They also preferred to have a negative profit shown rather than the accounting notation for loss. We discussed ways to change the question to make it clearer but no solutions were reached. Instead we considered inserting a statement prior to the question. The result is shown above and they agreed to "Accept with Revisions" to the text and the table. On the rating form, all students indicated agreement with the recommendation.

## APPENDIX G (continued)

### Question 11

*Alignment:* HSS-ID.7. Interpret the slope (rate of change) and the intercept (constant term) of a linear model in the context of the data.
*Key: C*
*Test: SAT*
*Calculator usage: NO CALC*
*Rigor: Application*
*Difficulty: Medium*

Jan is a lawyer who is researching legal cases related to water pollution in her area. Her initial search includes cases submitted from 1965 through 2000. The equation $f(x) = 11.5x + 14$ can be used to estimate the number of cases, $f(x)$, where $x$ is the number of years since 1964. What is the meaning of the constant 11.5 in this equation?

A. The number of cases in 1965.
B. The number of cases in 2000.
C. The increase in cases per year.
D. The increase in cases from 1965 to 2000.

*Accept As Is*

"I've had this problem before."

There was virtually no discussion about this question.
They agreed to "Accept As is". On the rating form, all students indicated agreement with the recommendation.

## APPENDIX G (continued)

### Question 12

*Alignment:* HSF-LE.2. Construct linear and exponential functions, including arithmetic and geometric sequences, given a graph, a description of a relationship, or two input-output pairs (include reading these from a table).
*Correct answer: 1/17 or .059*
*Test: SAT*
*Calculator usage: CALC*
*Rigor: Application*
*Difficulty: Hard*

*Accept with Revisions*

According to a census, the population of a town doubled approximately once every 17 years since 1950. If the equation $P = P_0 2^{kt}$, where $P_0$ is the population of the town in 1950, will be used to model the population of the town $t$ years after 1950, what should the value of $k$ be?

"I said that if you put once every 17 years, they'd be more inclined to put that"

There was virtually no discussion about this question. A suggestion was made to insert "once" before "every 17 years" and there was agreement with the change. Everyone indicated that they were able to solve the problem and they agreed to "Accept with Revisions" to the text. On the rating form, all students indicated agreement with the recommendation.

## APPENDIX G (continued)

### Question 13

$$\frac{1}{x} + \frac{2}{x} = \frac{1}{5}$$

Anise needs to complete a printing job and she will be using both printers in her office for it. One of the printers is twice as fast as the other and together they can finish the job in 5 hours. Anise wants to figure out how long would have taken her to finish the job if she had used the slower printer only. For this, she writes the equation above. What does the expression $\frac{1}{x}$ in this equation represent?

A. The time it takes the slower printer to finish the printing job.
B. The portion of job that the slower printer would complete in one hour.
C. The portion of job that the faster printer would complete in two hours.
D. The time it takes the slower printer to complete 1/5 of the printing job.

*Reject*

H = -16t^2 + 32t + 9

"Um, I thought that it was pretty clear, but um where it says she wants to know how long it would have taken her to finish the job if she had used the slower printer only. Like it doesn't really involve the question anyway afterwards. Like the question ends up asking what a value in that equation represents. So I thought it was kind of misleading. If it were taken out it would be very clear what it was asking. Yeah, cause I was originally thinking if you just didn't have the one fifth thing because it's only one of the hours, um you could do like x plus two x equals five and you wouldn't have to delete that section but I guess. I feel like it's better if you just delete that and keep the equation the same. That makes sense."
"Christina: I was just struggling between A and B when I was trying to understand it."
"I think it's pretty easy to get wrong."
"I just feel like some people won't make the connection between the one fifth and the five hours. They're not going to put it together."
"Yeah, cause when I think about it being on the no calculator portion, it's one of those questions that will probably be more time consuming. Like all the questions on that section are ones you have to think about because you don't have a calculator."
"Yeah, the fraction thing."
"Nobody likes fractions"
"I feel like they have to find the fraction in the word and they just can't. I would say change it to quadratic."

Almost all of the students struggled to solve this question. The one student who did solve it described why they thought the question was ok, and other students described why they were confused or that it would take a long time to analyze. We talked about the amount of time it would take and I provided examples of other questions that assess this alignment using quadratic or exponential models. Then, they shared concerns about how difficult fractions are. They decided that, due to the time required and the difficulty of the fractions to "Reject". On the rating form, the students were generally in agreement that rejecting this question and using a quadratic model would be better. However, 3 students said they would have also been fine keeping the question as is.

## APPENDIX G (continued)

### Question 14

The expression $\frac{1}{\sqrt{2.5} - \sqrt{0.4}}$ is equal to which of the following?

E. $\frac{1}{0.3}$

F. $\frac{1}{\sqrt{3}}$

G. $\frac{3}{\sqrt{10}}$ *Reject*

H. $\frac{\sqrt{10}}{3}$

"I think you need to give us more help on trying to find it."
"And then um, yeah I'm trying to remember back to last year when really did all these rules and like which ones are negatives and which are in the denominator. I guess it's not a trouble, you just need to know a couple rules like adding and subtracting powers and the rules. I guess there are like three different things you have to remember at the same time. It's not, since it is a hard difficulty, it's alright."
"I think it's a good hard one but at the same time, I am just blanking."
"I personally didn't like this question."
"I just don't know how to do it."
"I don't want to have bias and reject this just because I didn't know how to do it. But if there's somebody who knows how to do it."
"But this is basis math. Like basic math that you should know how to do."
"I really would not put this on the calculator portion. You just put this on your calculator and that would be pointless. I wouldn't reject this. Like I said before about another one. If I was, right now, if I was taking the SAT right now, I would have been studying stuff from a while ago. I mean this is, these rules, since they are kind of confusing, um like if you don't remember them, I would go back and study that myself um. So I think it should still be on here."
"I don't know, I don't know."
"I feel like if I got to a college calculator class and this was on like … I would just look it up and do it.
"It's not that I'm too stupid and I can't figure it out. I just don't know the rules on the back of my hand."
"I'm not going to sit there and memorize every kind of rule for this situation.'

None of the students knew how to solve this question. They described possible rules that they could use to solve the question, but could not string steps together to solve. Several students concluded that it was a hard question for them, but that it was ok to include on the SAT. I described how the question could be solved and one student said, "With what you said on how they want us to solve it, I feel like it has nothing to do with what it's [the alignment] saying." Other students agreed that it did not match the alignment. They then quickly agreed to "Reject". On the rating form, all students indicated agreement with the recommendation.

## APPENDIX G (continued)

### Question 15

*Alignment:* HSA-REI.10. Understand that the graph of an equation in two variables is the set of all its solutions plotted in the coordinate plane, often forming a curve (which could be a line).
*Key: B*
*Test: SAT*
*Calculator usage: CALC*
*Rigor: Conceptual understanding*
*Difficulty: Medium*

*Accept As Is*

Which of the following could be the graph of $y = mx$ in the $xy$-plane, where m is negative?

"So why is that a calculator question?"
"Yeah, that's a problem."
"Um, this one is really easy like, it's asking where m is negative and there"
"If you don't know what a negative graph is then, yeah, right are you going to college, should you be going to college? That's like Algebra 1"
"I feel like when I took the SAT there were maybe one or two questions that were really easy and I flew through them. It was like a freebee point like they gave it to you because like well, here's one that you don't have to."
'I mean it's kind of like identifying the basics to make sure that people know what a negative means on a graph. Cause they could probably…parabola and the parabola. It's just making sure you're not dumb."
"Cause like how do people get that question wrong? What are teachers teaching students?"
"I guess leave it on the calculator section."
"That's true cause you gotta do a lot of work"
"If you wanted to make it slightly more difficult you could do a b in here and"
"I think when people see y equals m x plus b, they immediately think slope and linear, but this one not."
"It's like, it's obvious. It's an obvious question."

The students knew how to solve this question. The discussion centered around whether the question belongs on the calculator or the no calculator portion and the difficulty of the question. Students suggested that the difficulty could be increased by putting it on the no calculator portion or by adding in the y-intercept constant and asking about that constant in addition to the slope constant. I asked the students if answer choice b was correct even though it didn't pass through the origin and several dismissed the concern, with the last student saying, "It's like, it's obvious. It's an obvious question." After the discussion, they agreed to "Accept As is". On the rating form, all students indicated agreement with the recommendation.

## APPENDIX G (continued)

### Question 16

*Alignment:* HSA-APR 3. Identify zeros of polynomials when suitable factorizations are available, and use the zeros to construct a rough graph of the function defined by the polynomial.
*Key: A*
*Test: SAT*
*Calculator usage: CALC*
*Rigor: Application*
*Difficulty: Medium*

$$y = x^3 + 2x^2 + x$$

If the equation shown is graphed on the *xy*-plane, which of the following points will the curve pass through?

A. (0,0)
B. (1,0) *Accept As Is*
C. (2,0)
D. (3,0)

"I would definitely take my calculator and graph it."
"It's a reading question."
"Yeah, and find the one it went through."
"I feel like I've done these questions before."
"At first I was like, why is it on the calculator portion. Then I realized just plug it in. my thought process was to, you know, …. Then you can plug it in your calculator."
"I mean it's a little concerning that this is considered medium. If you have a bunch, not a bunch, depending on how many easy questions you have, this is not even considered easy, that might make the whole test too easy."

The students knew how to solve this question and we had a very brief discussion. The discussion centered around whether the question belongs on the calculator or the no calculator portion and the difficulty of the question. After the discussion, they agreed to "Accept As is". On the rating form, all students indicated agreement with the recommendation.

## APPENDIX H

### Table XXIV

PERCENTAGE OF CONTROL GROUP THEMES BY QUESTION NUMBER

| | Math content | | | | Construct- | Cognitive psychology | | | | No | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Question | Alignment | Clarity | Difficulty | Other | irrelevance | Multiple | Calculators | Other | Consequences | theme | Total |
| 1 | 1.57% | 0.39% | 0.39% | 2.49% | 0.00% | 0.52% | 0.00% | 1.31% | 0.00% | 0.92% | 7.61% |
| 2 | 0.52% | 0.92% | 0.66% | 1.71% | 0.00% | 0.39% | 0.52% | 0.79% | 1.18% | 1.84% | 8.53% |
| 3 | 0.79% | 0.00% | 0.66% | 3.28% | 0.92% | 0.52% | 1.97% | 2.10% | 0.66% | 2.10% | 12.99% |
| 4 | 1.31% | 0.13% | 0.52% | 3.28% | 0.39% | 0.00% | 1.57% | 0.13% | 1.71% | 1.84% | 10.89% |
| 5 | 0.39% | 2.23% | 0.39% | 9.97% | 1.18% | 0.00% | 0.00% | 0.92% | 0.52% | 3.41% | 19.03% |
| 6 | 0.66% | 0.79% | 0.26% | 0.66% | 0.79% | 0.00% | 0.00% | 0.00% | 0.00% | 0.66% | 3.81% |
| 7 | 0.79% | 0.52% | 0.79% | 4.07% | 1.05% | 0.00% | 0.00% | 0.00% | 0.26% | 1.71% | 9.19% |
| 8 | 0.00% | 0.00% | 0.92% | 3.67% | 0.13% | 0.52% | 0.13% | 0.92% | 0.26% | 1.71% | 8.27% |
| 9 | 0.00% | 0.39% | 1.05% | 1.84% | 0.52% | 0.39% | 0.26% | 0.39% | 0.52% | 0.79% | 6.17% |
| 10 | 0.26% | 1.31% | 0.66% | 6.82% | 0.26% | 0.66% | 0.92% | 0.00% | 0.13% | 2.49% | 13.52% |
| Total | 6.30% | 6.69% | 6.30% | 31.23% | 5.25% | 3.02% | 5.38% | 3.67% | 5.25% | 17.45% | 100.00% |

### Table XXV

PERCENTAGE OF EXPERIMENTAL GROUP THEMES BY QUESTION NUMBER

| | Math content | | | | Construct- | Cognitive psychology | | | | No | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Question | Alignment | Clarity | Difficulty | Other | irrelevance | Multiple | Calculators | Other | Consequences | theme | Total |
| 1 | 2.44% | 1.58% | 1.87% | 6.75% | 0.57% | 0.86% | 0.14% | 1.87% | 0.29% | 5.75% | 22.13% |
| 2 | 2.01% | 0.43% | 2.30% | 7.18% | 0.57% | 0.14% | 0.14% | 2.59% | 2.30% | 5.75% | 23.42% |
| 3 | 1.15% | 0.00% | 1.58% | 3.59% | 0.14% | 0.72% | 0.43% | 1.29% | 0.57% | 1.58% | 11.06% |
| 4 | 1.29% | 0.00% | 1.01% | 4.31% | 0.57% | 0.14% | 1.44% | 1.87% | 0.72% | 3.88% | 15.23% |
| 5 | 0.14% | 2.30% | 1.15% | 13.07% | 2.30% | 0.14% | 0.29% | 0.86% | 0.86% | 7.04% | 28.16% |
| Total | 7.04% | 4.31% | 7.90% | 34.91% | 4.17% | 2.01% | 2.44% | 8.48% | 4.74% | 23.99% | 100.00% |

**Table XXVI**

PROPORTION OF THEMES FOR THE FIRST FIVE QUESTIONS BY GROUP

| | Math content | | | | Construct- | Cognitive psychology | | | | No | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Group | Alignment | Clarity | Difficulty | Other | irrelevance | Multiple | Calculators | Other | Consequences | theme | Total |
| Control | 7.78% | 6.22% | 4.44% | 35.11% | 4.22% | 2.44% | 6.89% | 8.89% | 6.89% | 17.11% | 100.00% |
| Experimental | 7.04% | 4.31% | 7.90% | 34.91% | 4.17% | 2.01% | 2.44% | 8.48% | 4.74% | 23.99% | 100.00% |
| Control | 7.78% | 6.22% | 4.44% | 35.11% | 4.22% | 2.44% | 6.89% | 8.89% | 6.89% | 17.11% | 100.00% |

## APPENDIX I

**Consensus recommendations (The comments in italic on the right are the consensus recommendations based on each groups' review.)**

### Question 1

**Test takers**

If cos $\theta = 0.5$, which of the following is also equal to 0.5?

A. sin $\theta$
B. tan $\theta$
C. sin $(90 - \theta)$
D. tan $(90 - \theta)$

*Accept with revisions to difficulty level only*

**Control group**

If cos $\theta = 0.5$, which of the following is also equal to 0.5?

A. sin θ
B. tan θ
C. sin $(90 - \theta)$
D. tan $(90 - \theta)$

*Reject due to alignment with state standard*

**Experimental group**

Which of the following is equal to cos $\theta$ for all values of $\theta$?

A. $\cos\left(\theta + \frac{\pi}{2}\right)$
B. $\cos(\theta + \pi)$
C. $\cos\left(\theta + \frac{7\pi}{2}\right)$
D. $\cos(\theta + 4\pi)$

*Accept with revisions to question and answer choices*

## APPENDIX I (continued)

### Question 2

**Test takers**

The standard equation of a circle is $(x–h)^2 + (y–k)^2 = r^2$.
A circle has the equation $x^2 – 6x + y^2 + 4y = 8$. What is the radius of this circle?

*Accept with revisions to question*

A. $\sqrt{3}$
B. $2\sqrt{2}$
C. $\sqrt{13}$
D. $\sqrt{21}$

**Control group**

In the *xy*-plane, what is the radius of the circle with equation $x^2 – 6x + y^2 + 4y = 8$?

*Accept as is*

A. $\sqrt{3}$
B. $2\sqrt{2}$
C. $\sqrt{13}$
D. $\sqrt{21}$

**Experimental group**

The standard equation of a circle is $(x–h)^2 + (y–k)^2 = r^2$.
A circle has the equation $x^2 – 6x + y^2 + 4y = 8$. What is the radius of this circle?

*Accept with revisions to question*

A. $\sqrt{3}$
B. $2\sqrt{2}$
C. $\sqrt{13}$
D. $\sqrt{21}$

## APPENDIX I (continued)

### Question 3

**Test takers**

If the equation $x^2 - 25x + c = 0$ has only one real solution, what is the value of c?

A. 5

B. $\frac{25}{2}$

C. $\frac{125}{4}$

D. $\frac{625}{4}$

*Accept as is*

**Control group**

If the equation $x^2 - 6x + c = 0$ has only one real solution, what is the value of c?

A. $\sqrt{6}$

B. 3

C. $\frac{27}{4}$

D. 9

*Accept with revisions to question and answer choices and move question to no calculator portion*

**Experimental group**

If the equation $x^2 - 25x + c = 0$ has only one real solution, what is the value of c?

A. 5

B. $\frac{25}{2}$

C. $\frac{125}{4}$

D. $\frac{625}{4}$

*Accept as is*

## APPENDIX I (continued)

### Question 4

**Test takers**

If $f(x) = \frac{x(x-1)}{2}$, which of the following is true about $f(x-1)$?

*Accept with revisions to answer choices*

A. $f(x-1) = \frac{(x-1)(x-2)}{2}$

B. $f(x-1) = \frac{(x-1)(x-1)}{2}$

C. $f(x-1) = \frac{(x-2)}{2}$

D. $f(x-1) = \frac{(x-1)}{2}$

**Control group**

If $f(x) = \frac{x(x-1)}{2}$, which of the following is true about $f(x-1)$?

*Accept with revisions to answer choices and move question to no calculator portion*

A. $f(x-1) = \frac{x^2-3x+2}{2}$

B. $f(x-1) = \frac{x^2-2x+1}{2}$

C. $f(x-1) = \frac{(x-2)}{2}$

D. $f(x-1) = \frac{(x-1)}{2}$

**Experimental group**

If $f(x) = \frac{x(x-1)}{2}$, which of the following is true about $f(x-1)$?

*Accept with revisions to answer choices*

A. $f(x-1) = \frac{(x-1)(x-2)}{2}$

B. $f(x-1) = \frac{(x-1)(x-1)}{2}$

C. $f(x-1) = f(x) - f(1)$

D. $f(x-1) = \frac{f(x) - f(1)}{2}$

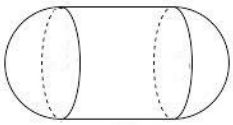
## APPENDIX I (continued)

### Question 5

**Test takers**

A gas tank manufacturer makes its tanks in the shape of a cylinder with hemispheres on both ends as shown.

*Accept with revisions to question*

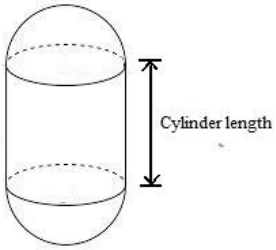


The diameter of the cylinder and hemisphere is always 40 inches, but the length of the cylinder can vary. If the manufacturer wants to make a tank that has a volume of 100,000 cubic inches, how long should the cylinder part of the tank measure, to the closest tenth of an inch?

**Control group**

A storage tank is in the shape of a cylinder with hemispheres on both ends as shown.

*Accept with revisions to question, and change to 4MC*




The diameters of the cylinder and hemispheres are 40 inches. If the volume of the tank is 100,000 cubic inches, which of the following is true?

A. The cylinder length is between 40 and 45 inches
B. The cylinder length is between 45 and 50 inches
C. The cylinder length is between 50 and 55 inches*
D. The cylinder length is between 55 and 60 inches

**Experimental group**

A gas tank manufacturer makes its tanks in the shape of a cylinder with hemispheres on both ends as shown.

*Accept with revisions to question*

The volume of the tank is 100,000 cubic inches. If the diameter of the cylinder is 40 inches, what is the length of the cylinder, to the nearest inch? (Use $\pi = 3.14$)

## APPENDIX I (continued)

### Question 6

**Test takers**

Aaron is staying at a hotel that charges $49.95 per night. The price does not include a tax of 8% on the price of the room and a one-time reservation fee of $1.00. Which of the following represents the price Aaron will pay if $x$ is the number of nights he will spend at the hotel?

A. $(49.95 + 0.08x) + 1$
B. $1.08(49.95x) + 1$
C. $1.08(49.95x + 1)$
D. $1.08(49.95 + 1)x$

*Accept as is*

**Control group**

Aaron is staying at a hotel that charges $49.95 per night. The price does not include a tax of 8% on the price of the room and a one-time reservation fee of $1.00. Which of the following represents the price Aaron will pay if $x$ is the number of nights he will spend at the hotel?

A. $(49.95 + 0.08x) + 1$
B. $1.08(49.95x) + 1$
C. $1.08(49.95x + 1)$
D. $1.08(49.95 + 1)x$

*Reject due to ambiguity and unfamiliarity of the context for students*

## APPENDIX I (continued)

### Question 7

**Test takers**

The acceleration, $a$, of a car can be found with the formula $a = (\frac{v_1 + v_2}{2})t$, where $v_1$ and $v_2$ are the initial and final speeds, respectively, and $t$ is the amount of time that elapsed between $v_1$ and $v_2$. Which equation represents the time elapsed in terms of the other variables?

*Accept with revisions to question*

A. $t = 2a(v_1 + v_2)$

B. $t = a\left(\frac{v_1 + v_2}{2}\right)$

C. $t = \frac{2a}{v_1 + v_2}$

D. $t = \frac{v_1 + v_2}{2a}$

**Control group**

The acceleration, $a$, of the car during this time can be found with the formula $a = \frac{v_2 - v_1}{t}$, where $v_1$ and $v_2$ are the initial and final speeds, respectively, and $t$ is the amount of time elapsed between $v_1$ and $v_2$. Which equation represents the time elapsed in terms of the other variables?

*Accept with revisions to question and answer choices*

A. $t = a(v_2 - v_1)$

B. $t = \frac{a(v_2 - v_1)}{2}$

C. $t = \frac{v_2 - v_1}{a}$

D. $t = \frac{2(v_2 - v_1)}{a}$

### Question 8

**Test takers**

If $-1 < -3t + 1 < \frac{1}{4}$, what is one possible value of $t$?

*Accept with Revisions*

**Control group**

If $-5 < -3t + 1 < -4$, what is one possible value of $9t - 3$?

*Accept with Revisions*

## APPENDIX I (continued)

### Question 9

**Test takers**

Given the system of equations:

$2x + 3y = 27$

$y = -5x + 12$

What is the $x$-value of solution $(x, y)$ to the system of equations above?

*Accept As Is*

**Control group**

$2x + 3y = 27$

$y = -5x + 12$

What is the $x$-value of the solution $(x, y)$ to the given system of equations?

*Accept with revisions to question*

## APPENDIX I (continued)

### Question 10

**Test takers**

Gene uses the following table to estimate the profit made based on the number of people that dine in his restaurant each day.

*Accept with revisions to question*

| Number of People | Estimated Profit |
|---|---|
| 15 | –$285 |
| 30 | $145 |
| 45 | $575 |
| 60 | $1005 |
| 75 | $1435 |

A linear function to represent this situation can be written in the form $p(x) = mx + b$, where m and b are constants. Which linear function represents the estimated profit, $p(x)$, based on the number of people, $x$, who dined in the restaurant?

A. $p(x) = \frac{86}{3}x - 715$

B. $p(x) = \frac{86}{3}x - 285$

C. $p(x) = 430x - 715$

D. $p(x) = 430x - 285$

**Control group**

The following table estimates the profit made based on the number of people that dine in the restaurant each day.

*Accept with revisions to question and answer choices*

| Number of People | Estimated Profit |
|---|---|
| 15 | $15 |
| 30 | $315 |
| 45 | $615 |
| 60 | $915 |
| 75 | $1215 |

Which linear function represents the estimated profit, $p(x)$, based on the number of people, $x$, who dine in the restaurant each day?

$p(x) = 15x + 15$

$p(x) = 15x - 285$

$p(x) = 20x + 15$

$p(x) = 20x - 285$

## APPENDIX I (continued)

### Question 11

**Test takers**

Jan is a lawyer who is researching legal cases related to water pollution in her area. Her initial search includes cases submitted from 1965 through 2000. The equation $f(x) = 11.5x + 14$ can be used to estimate the number of cases $f(x)$, where $x$ is the number of years since 1964. What is the meaning of the constant 11.5 in this equation?

A. The number of cases in 1965.
B. The number of cases in 2000.
C. The increase in cases per year.
D. The increase in cases from 1965 to 2000.

*Accept as is*

### Question 12

**Test takers**

According to a census, the population of a town doubled approximately once every 17 years since 1950. If the equation $P = P_0 2^{kt}$, where $P_0$ is the population of the town in 1950, will be used to model the population of the town $t$ years after 1950, what should the value of $k$ be?

*Accept with revisions to question*

### Question 13

**Test takers**

$$\frac{1}{x} + \frac{2}{x} = \frac{1}{5}$$

Anise needs to complete a printing job and she will be using both printers in her office for it. One of the printers is twice as fast as the other and together they can finish the job in 5 hours. Anise wants to figure out how long would have taken her to finish the job if she had used the slower printer only. For this, she writes the equation above. What does the expression $\frac{1}{x}$ in this equation represent?

A. The time it takes the slower printer to finish the printing job.
B. The portion of job that the slower printer would complete in one hour.
C. The portion of job that the faster printer would complete in two hours.
D. The time it takes the slower printer to complete 1/5 of the printing job.

*Reject due to the amount of time required to solve and high difficulty level of the fractions.*

## APPENDIX I (continued)

### Question 14

**Test takers**

The expression $\frac{1}{\sqrt{2.5} - \sqrt{0.4}}$ is equal to which of the following?

A. $\frac{1}{0.3}$

B. $\frac{1}{\sqrt{3}}$

C. $\frac{3}{\sqrt{10}}$

D. $\frac{\sqrt{10}}{3}$

*Reject due to misalignment with the state standard*

### Question 15

**Test takers**

Which of the following could be the graph of $y = mx$ in the $xy$-plane, where m is negative?

*Accept as is*

### Question 16

**Test takers**

$$y = x^3 + 2x^2 + x$$

If the equation shown is graphed on the $xy$-plane, which of the following points will the curve pass through?

A. (0,0)
B. (1,0)
C. (2,0)
D. (3,0)

*Accept as is*

## CITED LITERATURE


ACT. (2014). The ACT technical manual. Retrieved from https://www.act.org/content/dam/act/unsecured/documents/ACT_Technical_Manual.pdf

AERA, APA, & NCME. (2014). *Standards for educational and psychological testing* (5th ed.). Washington, DC: American Educational Research Association.

Aiken, L. R. (1980). Content validity and reliability of single items or questionnaires. *Educational and Psychological Measurement, 40*(4), 955-959. doi:10.1177/001316448004000419

Aiken, L. R. (1985). Three coefficients for analyzing the reliability and validity of ratings. *Educational and Psychological Measurement, 45*(1), 131-142. doi:10.1177/0013164485451012

Bandura, A. (1986). The explanatory and predictive scope of self-efficacy theory. *Journal of social and clinical psychology*, *4*(3), 359-373.

Bandura, A. (1997). *Self-efficacy: The exercise of control*. New York: W. H. Freeman.

Bandura, A. (2006). Guide for constructing self-efficacy scales. In F. Pajares & T. Urdan (Eds.), *Self-efficacy beliefs of adolescents* (pp. 307-337). Greenwich, CT: Information Age Publishing, Inc.

Begeny, J. C., Eckert, T. L., Montarello, S. A., & Storie, M. S. (2008). Teachers' perceptions of students' reading abilities: An examination of the relationship between teachers' judgments and students' performance across a continuum of rating methods. *School Psychology Quarterly, 23*(1), 43-55. doi:10.1037/1045-3830.23.1.43

Berk, R. A. (1990). Importance of expert judgment in content-related validity evidence. *Western Journal of Nursing Research, 12*(5), 659-671. doi:10.1177/019394599001200507

Bloom, B. S., Englehart, M., Furst, E. J., Hill, W. H., & Krathwohl, D. R. (1956). *Taxonomy of educational objectives: Handbook I cognitive domain*. New York: David McKay.

Bradley, K. D., Peabody, M. R., Akers, K. S., & Knutson, N. (2015). Rating scales in survey research: using the Rasch model to illustrate the middle category measurement flaw. *Survey Practice*, *8*(2), 1-14. doi:10.29115/SP-2015-0001

Brennan, R. L. (2006). *Educational measurement* (4th ed.). New York, NY: Macmillan.

Bridge, P. D., Musial, J., Frank, R., Roe, T., & Sawilowsky, S. (2003). Measurement practices: methods for developing content-valid student examinations. *Med Teach, 25*(4), 414-421. doi:10.1080/0142159031000100337

Camilli, G. (2006). Test fairness. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 221-256). Westport, CT: Praeger Publishers.

Clinedinst, M., Koranteng, A. M., & Nicola, T. (2016). *2015 state of college admission*. Retrieved from https://indd.adobe.com/view/c555ca95-5bef-44f6-9a9b-6325942ff7cb

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement, 20*(1), 37-46.

Coladarci, T. (1986). Accuracy of teacher judgments of student responses to standardized test items. *Journal of Educational Psychology, 78*(2), 141-146.

College Board. (2017). SAT technical manual. Retrieved from https://collegereadiness.collegeboard.org/pdf/sat-suite-assessments-technical-manual.pdf

Creswell, J. W. (2002). *Educational research: Planning, conducting, and evaluating quantitative*. New Jersey: Prentice Hall Upper Saddle River.

Crocker, L. (2003). Teaching for the test: Validity, fairness, and moral action. *Educational Measurement: Issues and Practice, 22*(3), 5-11.

D'Agostino, J., Karpinski, A., & Welsh, M. (2011). A method to examine content domain structures. *International Journal of Testing, 11*(4), 295-307. doi:10.1080/15305058.2011.570885

Delgado-Rico, E., Carretero-Dios, H., & Ruch, W. (2012). Content validity evidences in test development: An applied perspective. *International journal of clinical and health psychology, 12*(3), 449.

DeMars, C. E., & Erwin, T. D. (2004). Scoring neutral or unsure on an identity development instrument for higher education. *Research in higher education, 45*(1), 83-95.

Downing, S. M., & Haladyna, T. M. (1997). Test item development: Validity evidence from quality assurance procedures. *Applied Measurement in Education, 10*(1), 61-82.

Eckert, T. L., Dunn, E. K., Codding, R. S., Begeny, J. C., & Kleinmann, A. E. (2006). Assessment of mathematics and reading performance: An examination of the correspondence between direct assessment of student performance and teacher report. *Psychology in the Schools, 43*(3), 247-265. doi:10.1002/pits.20147

Educational Testing Service. (2009). ETS international principles for fairness review of assessments. Retrieved from https://www.ets.org/s/about/pdf/standards.pdf

Eugenio, B. D., & Glass, M. (2004). The kappa statistic: A second look. *Computational linguistics, 30*(1), 95-101.

Feinberg, A. B., & Shapiro, E. S. (2009). Teacher accuracy: An examination of teacher-based judgments of students' reading with differing achievement levels. *The Journal of Educational Research, 102*(6), 453-462.

Fink, A., Kosecoff, J., Chassin, M., & Brook, R. H. (1984). Consensus methods: Characteristics and guidelines for use. *American journal of public health, 74*(9), 979-983.

Gallagher, A. M., & De Lisi, R. (1994). Gender differences in Scholastic Aptitude Test: Mathematics problem solving among high-ability students. *Journal of Educational Psychology, 86*(2), 204-211.

Gierl, M. J. (1997). Comparing cognitive representations of test developers and students on a mathematics test With Bloom's taxonomy. *The Journal of Educational Research, 91*(1), 26-32. doi:10.1080/00220679709597517

Grant, J. S., & Davis, L. L. (1997). Selection and use of content experts for instrument development. *Res Nurs Health, 20*(3), 269-274.

Gulikers, J., Biemans, H., & Mulder, M. (2009). Developer, teacher, student and employer evaluations of competence-based assessment quality. *Studies in Educational Evaluation, 35*(2-3), 110-119. doi:10.1016/j.stueduc.2009.05.002

Haertel, E. H., & Lorie, W. A. (2004). Validating standards-based test score interpretations. *Measurement, 2*(2), 61-103.

Haladyna, T., & Roid, G. (1981). The role of instructional sensitivity in the empirical review of criterion-referenced test items. *Journal of Educational Measurement, 18*(1), 39-53.

Haladyna, T. M., & Downing, S. M. (2004). Construct-irrelevant variance in high-stakes testing. *Educational Measurement: Issues and Practice, 23*(1), 17-27.

Hamilton, L. S., Nussbaum, E. M., & Snow, R. E. (1997). Interview procedures for validating science assessments. *Applied Measurement in Education, 10*(2), 181-200.

Harpe, S. E. (2015). How to analyze Likert and other rating scale data. *Currents in Pharmacy Teaching and Learning, 7*(6), 836-850. doi:10.1016/j.cptl.2015.08.001

Herman, J. L., Webb, N. M., & Zuniga, S. (2003). *Alignment and college admissions: The match of expectations, assessments, and educator perspectives.* Paper presented at the annual meeting of the American Education Research Association, New Orleans, LA.

Hess, K. K., Jones, B. S., Carlock, D., & Walkup, J. R. (2009). Cognitive rigor: Blending the strengths of Bloom's taxonomy and Webb's depth of knowledge to enhance classroom-level processes. [ED517804]. Retrieved from http://eric.ed.gov/

Hoge, R. D., & Coladarci, T. (1989). Teacher-based judgments of academic achievement: A review of literature. *Review of Educational Research, 59*(3), 297-313.

Hopkins, K. D., George, C. A., & Williams, D. D. (1985). The concurrent validity of standardized achievement tests by content area using teachers' ratings as criteria. *Journal of Educational Measurement, 22*(3), 177-182.

Huitema, B. (2011). *The analysis of covariance and alternatives: Statistical methods for experiments, quasi-experiments, and single-case studies*. Hoboken, NJ: John Wiley & Sons.

Johnstone, C., Altman, J., Thurlow, M., & Moore, M. (2006). Universal design online manual. Retrieved from https://nceo.info/Resources/publications/UDmanual/default.html

Kane, M. T. (2006). Validation. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 17-64). Westport, CT: Praeger Publishers.

Kane, M. T. (2013). Validating the interpretations and uses of test scores. *Journal of Educational Measurement, 50*(1), 1-73.

Kane, M. T. (2016). Validation strategies: Delineating and validating proposed interpretations and uses of test scores. In S. Lane, M. R. Raymond, & T. M. Haladyna (Eds.), *Handbook of test development* (2nd ed., pp. 64-80). New York, NY: Routledge.

Katz, I. R., Bennett, R. E., & Berger, A. E. (2000). Effects of response format on difficulty of SAT-mathematics items: It's not the strategy. *Journal of Educational Measurement, 37*(1), 39-57.

Kerlinger, F., & Lee, H. (2000). *Foundations of behavioral research* (4th ed.). New York: International Thomson Publishing.

Koo, T. K., & Li, M. Y. (2016). A guideline of selecting and reporting intraclass correlation coefficients for reliability research. *Journal of chiropractic medicine, 15*(2), 155-163.

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *biometrics, 33*(1), 159-174.

Lane, S., Raymond, M. R., Haladyna, T. M., & Downing, S. M. (2016). Test development process. In S. Lane, M. R. Raymond, & T. M. Haladyna (Eds.), *Handbook of test development* (2nd ed., pp. 3-18). New York, NY: Routledge.

Leighton, J. P., & Gokiert, R. J. (2005). *The cognitive effect of test item features: Informing item generation by identifying construct irrelevant variance*. Paper presented at the Annual Meeting of the National Council on Measurement in Education, Montreal, Quebec, Canada.

Lennon, R. T. (1956). Assumptions underlying the use of content validity. *Educational and Psychological Measurement, 16*(3), 294-304. doi:10.1177/001316445601600303

Li, X., & Sireci, S. G. (2013). A new method for analyzing content validity data using multidimensional scaling. *Educational and Psychological Measurement, 73*(3), 365-385. doi:10.1177/0013164412473825

Linn, R. L. (1989). *Educational measurement* (3rd ed.). New York, NY: Macmillan.

Marzano, R. J. (2001). *Designing a new taxonomy of educational objectives*. Thousand Oaks, CA: Corwin Press, Inc., A Sage Publications Company.

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13-103). New York: MacMillan.

Michigan Department of Education. (2010). Michigan K-12 standards for mathematics. Retrieved from http://www.michigan.gov/documents/mde/K-12_MI_Math_Standards_REV_470033_7_550413_7.pdf

Mislevy, R. J. (2006). Cognitive Psychology and Educational Assessment. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 257-306). Westport, CT: Praeger Publishers.

Morris, D. B., Usher, E. L., & Chen, J. A. (2016). Reconceptualizing the sources of teaching self-efficacy: A critical review of emerging literature. *Educational Psychology Review*, 1-39. doi:10.1007/s10648-016-9378-y

Muijs, D., & Reynolds, D. (2015). Teachers' beliefs and behaviors: what really matters? *The Journal of Classroom Interaction, 50*(1), 25.

Nunnally, J. C. (1978). *Psychometric theory* (2nd ed.). Hillsdale, NJ: Mcgraw-Hill.

Pellegrino, J. W., Chudowsky, N., & Glaser, R. (2001). *Knowing what students know: The science and design of educational assessment*. Washington, D.C.: National Academies Press.

Polin, L., & Baker, E. L. (1979, April). *Qualitative analysis of test item attributes for domain-referenced content validity judgments.* Paper presented at the Annual Meeting of the American Educational Research Association, San Francisco, CA.

Rovinelli, R. J., & Hambleton, R. K. (1976). On the use of content specialists in the assessment of criterion-referenced test item validity. *Dutch Journal of Educational Research, 2*, 49–60.

Rubio, D. M., Berg-Weger, M., Tebb, S. S., Lee, E. S., & Rauch, S. (2003). Objectifying content validity: Conducting a content validity study in social work research. *Social work research, 27*(2), 94-104.

Ryan, J. J. (1968). Teacher judgments of test item properties. *Journal of Educational Measurement, 5*(4), 301-306.

Ryan, K. E. (2002). Assessment validation in the context of high-stakes assessment. *Educational Measurement: Issues and Practice, 21*(1), 7-15.

Ryan, K. E., Ryan, A. M., Arbuthnot, K., & Samuels, M. (2007). Students' motivation for standardized math exams. *Educational Researcher, 36*(1), 5-13. doi:10.3102/0013189X06298001

Saldaña, J. (2009). *The coding manual for qualitative researchers*. London: Sage.

Schilling, L. S., Dixon, J. K., Knafl, K. A., Grey, M., Ives, B., & Lynn, M. R. (2007). Determining content validity of a self-report instrument for adolescents using a heterogeneous expert panel. *Nursing Research, 56*(5), 361-366.

Schmeiser, C. B., & Welch, C. J. (2006). Test development. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 307-353). Westport, CT: Praeger Publishers.

Serder, M., & Jakobsson, A. (2014). "Why bother so incredibly much?": Student perspectives on PISA science assignments. *Cultural Studies of Science Education, 10*(3), 833-853. doi:10.1007/s11422-013-9550-3

Sireci, S. G. (1998). Gathering and analyzing content validity data. *Educational Assessment, 5*(4), 299-321.

Sireci, S. G., & Geisinger, K. F. (1992). Analyzing test content using cluster analysis and multidimensional scaling. *Applied Psychological Measurement, 16*(1), 17-31. doi:10.1177/014662169201600102

Sireci, S. G., & Geisinger, K. F. (1995). Using subject-matter experts to assess content representation: An MDS analysis. *Applied Psychological Measurement, 19*(3), 241-255. doi:10.1177/014662169501900303

Stewart, J. L., Lynn, M. R., & Mishel, M. H. (2005). Evaluating content validity for children's self-report instruments using children as content experts. *Nursing Research, 54*(6), 414-418.

Strauss, A., & Corbin, J. (1998). *Basics of qualitative research: Procedures and techniques for developing grounded theory*. Thousand Oaks, CA: Sage.

Trapp, W. J. (2015). *The impact of test taker feedback on subject matter expert review.* Paper presented at the Annual Meeting of the American Educational Research Association, Chicago, IL.

Tschannen-Moran, M., & Hoy, A. W. (2001). Teacher efficacy: Capturing an elusive construct. *Teaching and teacher Education, 17*(7), 783-805.

Vogt, D. S., King, D. W., & King, L. A. (2004). Focus groups in psychological assessment: Enhancing content validity by consulting members of the target population. *Psychol Assess, 16*(3), 231-243. doi:10.1037/1040-3590.16.3.231

Webb, N. L. (1997). Criteria for Alignment of Expectations and Assessments in Mathematics and Science Education. Research Monograph No. 6. Retrieved from http://facstaff.wceruw.org/normw/WEBBMonograph6criteria.pdf

Wolfe, E. W., & Smith, J. E. (2007). Instrument development tools and activities for measure validation using Rasch models: Part II--validation activities. *Journal of applied measurement, 8*(2), 204-234.

Wynd, C. A., & Schaefer, M. A. (2002). The osteoporosis risk assessment tool: Establishing content validity through a panel of experts. *Applied Nursing Research, 15*(3), 184-188. doi:10.1053/apnr.2002.34243

Yaghmale, F. (2009). Content validity and its estimation. *Journal of Medical Education, 3*(1), 25-27.

Zieky, M. (2016). Developing fair tests. In S. Lane, M. R. Raymond, & T. M. Haladyna (Eds.), *Handbook of test development* (2nd ed., pp. 81-99). New York, NY: Routledge.

Zwick, R. (2006). Higher education admissions testing. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 647-680). Westport, CT: Praeger Publishers.

# VITA

| | |
|---|---|
| NAME | William James Trapp |
| EDUCATION: | B.S., Mathematics, Western Illinois University, Macomb, Illinois, 1997 |
| | M.S., Educational Technology, Research, and Assessment, Northern Illinois University, Dekalb, Illinois, 2007 |
| | Ph.D., Educational Psychology, University of Illinois at Chicago, Chicago, Illinois, 2018 |
| TEACHING: | Department of Educational Technology, Research, and Assessment, Northern Illinois University, 2009 – 2011 |
| | Mathematics and Physics, Wauconda High School, 1998 – 2001 |
| EXPERIENCE: | Executive Director, Math and Science Assessment, College Board (2013 -present) |
| | Director of Product Development, Riverside Publishing (2001 – 2013) |
| PROFESSIONAL MEMBERSHIP: | National Council of the Teachers of Mathematics |
| PROJECTS: | Trapp, W.J. (2007). *Criteria to evaluate interpretive guides for criterion-referenced rests.* Unpublished master's project, Northern Illinois University, Dekalb, IL. |
| PRESENTATIONS: | Trapp, W.J. (2017, November). *Reflections on the PSAT/NMSQT and the SAT from the College Board Math Test Development Committee*. Session at the regional meeting of the National Council of the Teachers of Mathematics, Chicago, IL. |
| | Trapp, W.J., & Wilcox, L. (2017, April). *Reflections from the Redesigned SAT Test Development Committee*. Session at the annual meeting of the National Council of the Teachers of Mathematics, San Antonio, TX. |
| | Trapp, W.J. (2015, April). *The impact of test taker feedback on subject matter expert review.* Paper presented at the annual meeting of the American Educational Research Association, Chicago, IL. |
| | Trapp, W.J. (2013, November). *Gathering content validity evidence from test takers – asking the right questions*. Paper presented at the annual meeting of the Midwestern Educational Research Association, Evanston, IL. |

Trapp, W.J. (2012, November). *The effect of standard setting judge variance on assessment cut scores*. Paper presented at the annual meeting of the Midwestern Educational Research Association, Evanston, IL.

Rispoli, M., Hadley, P.A., Holt, J,K, & Trapp, W.J. (2010, June). *Sequence and system in the development of tense and agreement.* Poster presented at the annual Symposium for Research in Child Language Disorders, Madison, WI.

Trapp, W.J. (2007, October). *Sources of state department guidance on the topic of data driven decision making.* Paper presented at the annual meeting of the Midwestern Educational Research Association, St. Louis, MO.

Trapp, W.J. (2007, October). *Checkerboard graphical displays for longitudinal, multivariate data*. Paper presented at the annual meeting of the Midwestern Educational Research Association, St. Louis, MO.