Research Report

Anchored Graphical Representations: A Graphical Alternative to Traditional Just Qualified Candidate Descriptors for Licensure Tests

ETS RR–18-40

Priya Kannan
Richard J. Tannenbaum
Delano Hebert

December 2018

Discover this journal online at
Wiley Online Library
wileyonlinelibrary.com

ETS Research Report Series

EIGNOR EXECUTIVE EDITOR

James Carlson
Principal Psychometrician

ASSOCIATE EDITORS

Beata Beigman Klebanov
Senior Research Scientist

Heather Buzick
Senior Research Scientist

Brent Bridgeman
Distinguished Presidential Appointee

Keelan Evanini
Research Director

Marna Golub-Smith
Principal Psychometrician

Shelby Haberman
Consultant

Anastassia Loukina
Research Scientist

John Mazzeo
Distinguished Presidential Appointee

Donald Powers
Principal Research Scientist

Gautam Puhan
Principal Psychometrician

John Sabatini
Managing Principal Research Scientist

Elizabeth Stone
Research Scientist

Rebecca Zwick
Distinguished Presidential Appointee

PRODUCTION EDITORS

Kim Fryer
Manager, Editing Services

Ayleen Gontz
Senior Editor

Since its 1947 founding, ETS has conducted and disseminated scientific research to support its products and services, and to advance the measurement and education fields. In keeping with these goals, ETS is committed to making its research freely available to the professional community and to the general public. Published accounts of ETS research, including papers in the ETS Research Report series, undergo a formal peer-review process by ETS staff to ensure that they meet established scientific and professional standards. All such ETS-conducted peer reviews are in addition to any reviews that outside organizations may provide as part of their own publication processes. Peer review notwithstanding, the positions expressed in the ETS Research Report series and other published accounts of ETS research are those of the authors and not necessarily those of the Officers and Trustees of Educational Testing Service.

The Daniel Eignor Editorship is named in honor of Dr. Daniel R. Eignor, who from 2001 until 2011 served the Research and Development division as Editor for the ETS Research Report series. The Eignor Editorship has been created to recognize the pivotal leadership role that Dr. Eignor played in the research publication process at ETS.

RESEARCH REPORT

Anchored Graphical Representations: A Graphical Alternative to Traditional Just Qualified Candidate Descriptors for Licensure Tests

Priya Kannan, Richard J. Tannenbaum, & Delano Hebert

Educational Testing Service, Princeton, NJ

A well-constructed just qualified candidate (JQC) description is needed to arrive at a reasonable passing score for licensure tests. Traditionally, such descriptions consist of a list of knowledge and skill statements without sufficient context to internalize its intended meaning, allowing the standard-setting panelists to make idiosyncratic interpretations. In a series of studies, we evaluated an alternative JQC description called anchored graphical representation (AGR). AGRs are intended to provide both text and visual organizers by each tested domain to contextualize the meaning of the JQC. In Study 1, 22 mathematics educators participated in a mock standard-setting study and were randomly assigned to conditions in which they used either the AGR-based or the text-based definitions in making standard-setting judgments. In Study 2, 17 social studies educators were randomly assigned to either the AGR-based or text-based condition. While the overall passing score did not noticeably vary between the conditions, consistent with our expectations, results from both studies showed that the domain passing scores for the AGR groups reflected the relative importance of each test domain for the JQC and that the inter-panelist variability was smaller for the AGR groups. Collectively, our results indicate that use of the AGR resulted in standard-setting ratings that were more aligned with the panelists' differential expectations of JQCs' performance for each domain, and the results add to the procedural validity evidence supporting the reasonableness of standard-setting recommendations.

Keywords Standard setting; performance level descriptors; educator licensure

doi:10.1002/ets2.12228

Standard setting refers to a variety of systematic, judgment-based processes that identify a minimum test score that separates one level of performance (e.g., understanding, competence, expertise, or accomplishment) from another (Tannenbaum, 2011). Cizek (1993) defined standard setting as the "proper following of a prescribed, rational system of rules or procedures resulting in the assignment of a number to differentiate between two or more conceivable states or degrees of performance" (p. 100). Key to any standard-setting application is constructing operational definitions of the performance standards, which are then applied to differentiate performance levels by specific points on the test score scale (i.e., the cut scores). Each cut score represents the transition point between the highest acceptable score within a lower level and the lowest acceptable score within the adjacent higher level. Each level is operationally defined by a performance-level descriptor (PLD). For example, in a K–12 context, there are typically multiple PLDs, such as basic, proficient, and advanced. One cut score marks the transition from basic to proficient, and another marks the transition from proficient to advanced. In educator licensure testing, one cut score differentiates the pass status from the fail status. The focus of standard setting is on the minimum requirements for entrance into the higher of the two adjacent performance levels. In other words, the operational definition of the performance standard focuses on the test-relevant knowledge and/or skills that are minimally sufficient to reach the higher level. A test taker with just-sufficient knowledge and/or skills is variously referred to as a borderline test taker, a minimally competent candidate, a target test taker, or, as we will use, a just qualified candidate (JQC).

In educator licensure testing, a well-constructed descriptor of the JQC is needed to arrive at a reasonable passing score (Egan, Schneider, & Ferrara, 2012; Perie, 2008). When clearly defined and consistently used, a well-constructed descriptor will result in consistent cut score decisions. For example, Plake, Impara, and Irwin (1999) found that when judges (panelists) set cut scores on the same set of common items using the same target student definitions on two different occasions, they set essentially the same cut score both times. Some judges served on both panels, and Plake et al. conducted

Corresponding author: P. Kannan, E-mail: pkannan@ets.org

separate analyses for both repeating judges and one-time panelists and found that both groups set essentially the same cut scores on these items on each occasion. Similarly, across nine different teacher licensure assessments, Tannenbaum and Kannan (2015) showed that when two independent multistate panels (Tannenbaum, 2011) of educators used the same JQC definition, they provided consistent judgments that resulted in comparable cut scores between the panels. However, without a clearly defined descriptor, standard-setting panelists' ratings are more susceptible to idiosyncratic interpretations and irrelevant sources of variance (Hambleton, 2001). In two workshops, separated by a few months, Impara, Giraud, and Plake (2000) showed that when panelists were provided with different definitions of the target student, their cut score recommendations varied. Moreover, Impara et al. also found that the panelists were less consistent among themselves (i.e., higher variance) during the second workshop when compared to the first. Impara et al. contended that "cut scores are context-dependent and constructed based on the judges' interpretation of the definition provided and by the discussion" (p. 14) and thereby substantiated the critical importance of the target student (i.e., the JQC) definition in the standard-setting process.

Traditionally, these target test taker descriptors tend to take the form of a bulleted list of knowledge/skill statements (Cizek, 2012; Cizek & Bunch, 2007; Egan et al., 2012; Hambleton & Pitoniak, 2006; Perie, 2008). In educator licensure testing, in which one cut score is needed to differentiate pass from fail, a full-scope PLD is not often defined; rather, standard-setting panelists construct the JQC, focusing only on defining the minimally acceptable knowledge and/or skills to merit passing the test. However, a typical JQC description, consisting largely of knowledge and/or skill statements, may not provide sufficient context for the standard-setting panelist to internalize its intended meaning. For example, a text-only representation may not make it clear if more or less competency is expected of the JQC in different test domains. A single passing score on a licensure test often means that a compensatory decision model is employed—the passing score may be achieved by demonstrating more competence on certain domains than on others. It seems reasonable, therefore, that explicitly organizing the JQC description by content domain and in terms of contextualized expectations for competence in each domain may offer a clearer frame of reference against which to make standard-setting judgments.

The need for better contextualization may be a more relevant question for educator licensure testing, in which there is no progression of performance levels or expectations, as there is in K–12 testing. In a sense, the progression from basic to proficient to advanced offers a comparative framework for panelists in K–12 testing. Proficient, for example, is defined in absolute terms, as a stand-alone expectation, but also benefits somewhat from having the descriptors for basic and advanced. Other than, perhaps, operationally defining the deficit of knowledge and/or skills that justifies a fail status, which may not be a reasonable thing to do and may inadvertently establish an unintended bias, there is no frame of reference against which to help clarify the meaning of the JQC in standard setting for licensure tests.

Therefore the purpose of the current study was to investigate an alternative representation of a JQC for educator licensure standard-setting use—termed an anchored graphical representation (AGR). The concept of the AGR was informed by three interrelated areas of research: graphic organizers (e.g., Stull & Mayer, 2007), frame-of-reference (FOR) training (e.g., Bernardin & Buckley, 1981; Pulakos, 1984; Sulsky & Kline, 2007), and behaviorally anchored rating scales (BARS; e.g., Bernardin & Beatty, 1984; Chung & Khan, 2008; Pounder, 2000). These theoretical underpinnings are reviewed in the following sections.

Review of Literature

Graphic Organizers

According to Stull and Mayer (2007), graphic organizers consist "of spatial arrangements of words (or word groups) intended to represent the conceptual organization of text" (p. 180). Most people are familiar with these types of diagrams in textbooks, in which authors support readers' understanding by presenting information through multiple representations (the text and the supporting graphic organizer) of information. Graphic organizers include concept maps, hierarchies, matrices, and similar diagrams. Graphic organizers have been used to aid retention of new information (Diekhoff, Brown, & Dansereau, 1982), allowing viewers to see how concepts fit together (e.g., Hall, Hall, & Saling, 1999). Bauer and Johnson-Laird (1993) demonstrated how diagrams can help foster reasoning; compared with a text-based representation of a situation, people made faster and more valid inferences about the situation when presented with a diagrammatic representation that made alternative possibilities more explicit. Gobert and Clement (1999) compared students who created textual summaries to students who created graphical (diagrammatic) summaries of scientific text. They found that the textual summaries had more detail (more semantic information), whereas the graphical summaries enabled participants to recall more information. The participants in the graphical group also performed comparatively better on the posttest measure. These researchers reasoned that creating diagrams promotes a deeper understanding of the material and therefore aids retention.

In this study, we were influenced by this research on the use of graphic organizers and extended it to the use of a graphic organizer (the AGR) to represent domain-level expectations for a JQC using reference markers (anchors). These anchors were descriptions of what is expected of test takers at the top, middle, and bottom of the knowledge continuum for each test domain. The specific use of the anchor test takers in enhancing understanding was further informed by research in the areas of FOR training (e.g., Bernardin & Buckley, 1981; Pulakos, 1984; Sulsky & Kline, 2007) and BARS (e.g., Bernardin & Beatty, 1984; Chung & Khan, 2008; Pounder, 2000).

Frame-of-Reference Training and Behaviorally Anchored Rating Scales

FOR training and BARS are intended to improve the accuracy and consistency of raters' judgments by providing common or shared benchmarks against which judgments are made. These tools were initially conceived in the context of improving performance appraisal ratings, in which the goal was to reduce the likelihood of raters making idiosyncratic judgments of work performance (e.g., Bernardin & Buckley, 1981). In FOR training, raters first discuss the meaning of each job dimension (domain) of interest and the types of behaviors or critical incidents that are indicative of different levels of effectiveness on each specific dimension. They subsequently practice making ratings using the rating scales—by, for example, observing a prescored vignette of performance—provide rationales for their ratings, and receive feedback on accuracy (proximity to the prescore) of their ratings (Bernardin & Buckley, 1981; Pulakos, 1984; Woehr & Huffcutt, 1994). As noted by Sulsky and Kline (2007), "FOR training is designed to provide raters with a theory of performance for each performance dimension to be evaluated" (p. 122). Evidence from research (Day & Sulsky, 1995; Schleicher, Day, Mayes, & Riggio, 2002; Uggerslev & Sulsky, 2008) has suggested that FOR training improves rater accuracy.

BARS, initially proposed by Smith and Kendall (1963), were intended to improve the accuracy of performance ratings over that which had typically been observed using Likert-type rating scales. Instead of the simple descriptive adjectives used in Likert-type scales, which may lead to idiosyncratic ratings, BARS use graphic rating scales with behavioral descriptions illustrating various points on a vertical scale of performance (Bernardin & Beatty, 1984). Traditionally, for each job dimension, experts (e.g., job incumbents and supervisors) generate critical incidents or specific examples of more and less effective behavior pertaining to that dimension. Another group of experts (often the raters themselves) then retranslates these critical incidents back into their purported dimensions. Those incidents reliably retranslated are then rated by this same group of experts in terms of performance effectiveness using a scale (e.g., 1–5, 1–7). The averages (i.e., medians) of these effectiveness ratings are then used to associate each incident with a scale value (e.g., Cook, 1989; Pounder, 2000; Schwab, Heneman, & DeCotiis, 1975) and are placed on a vertical scale to develop the BARS (see Figure 1 for an example of a BARS). The purported benefit of using behavioral anchors is that they can more objectively delineate what is expected at different scale points within a dimension and help reduce subjectivity and personal bias (Stoskopf, Glik, Baker, Ciesla, & Cover, 1992) in raters' interpretations of those scale levels. It has also been shown that the BARS yield fewer leniency and halo errors and greater discriminant validity than various other types of unanchored rating scales (Campbell, Dunnette, Arvey, & Hellervick, 1973). Because the BARS technique was developed to facilitate a common frame of reference among raters, it seems particularly well suited for supporting standard-setting panelists' judgments.

From a standard-setting perspective, for licensure testing, the goal of developing a JQC definition is to facilitate all panelists developing a common understanding (frame of reference) of the knowledge and skill levels of a JQC such that they can distinguish successful from unsuccessful candidates. Moreover, as applied to multiple-choice, item-level judgments, the JQC definition should help all panelists come to similar conclusions about the probability with which a JQC can get an item correct. Therefore, in this study, we hypothesized that graphically representing the JQC with anchored frames of reference on a performance scale for each test dimension (based on a BARS-like framework) might be helpful for standard-setting panelists.

Anchored Graphical Representations of the Just Qualified Candidate

Inspired by the three interrelated areas of research described earlier, in this report, we propose an alternative method of defining the JQC—the AGR—that includes the use of both visual/graphic organizers and benchmarks (or anchors)

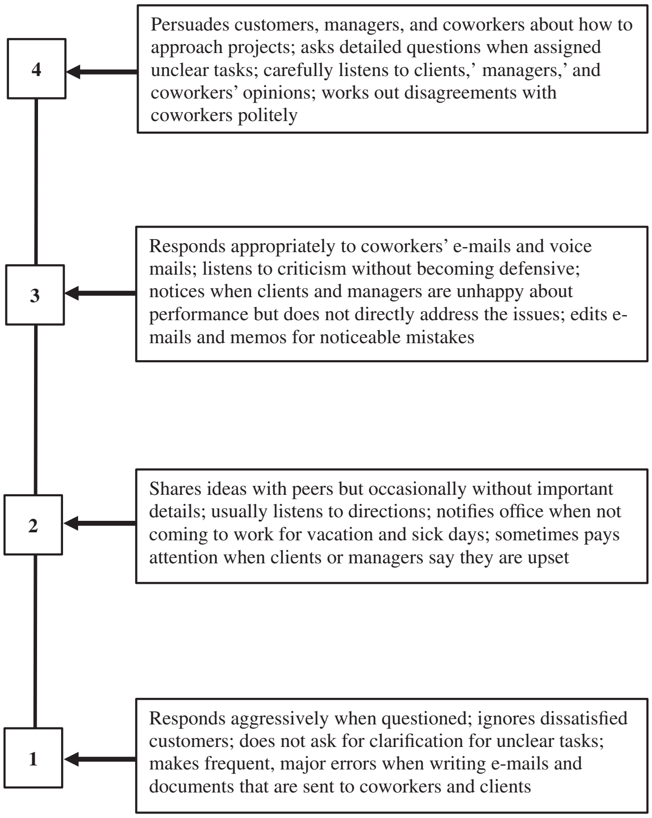
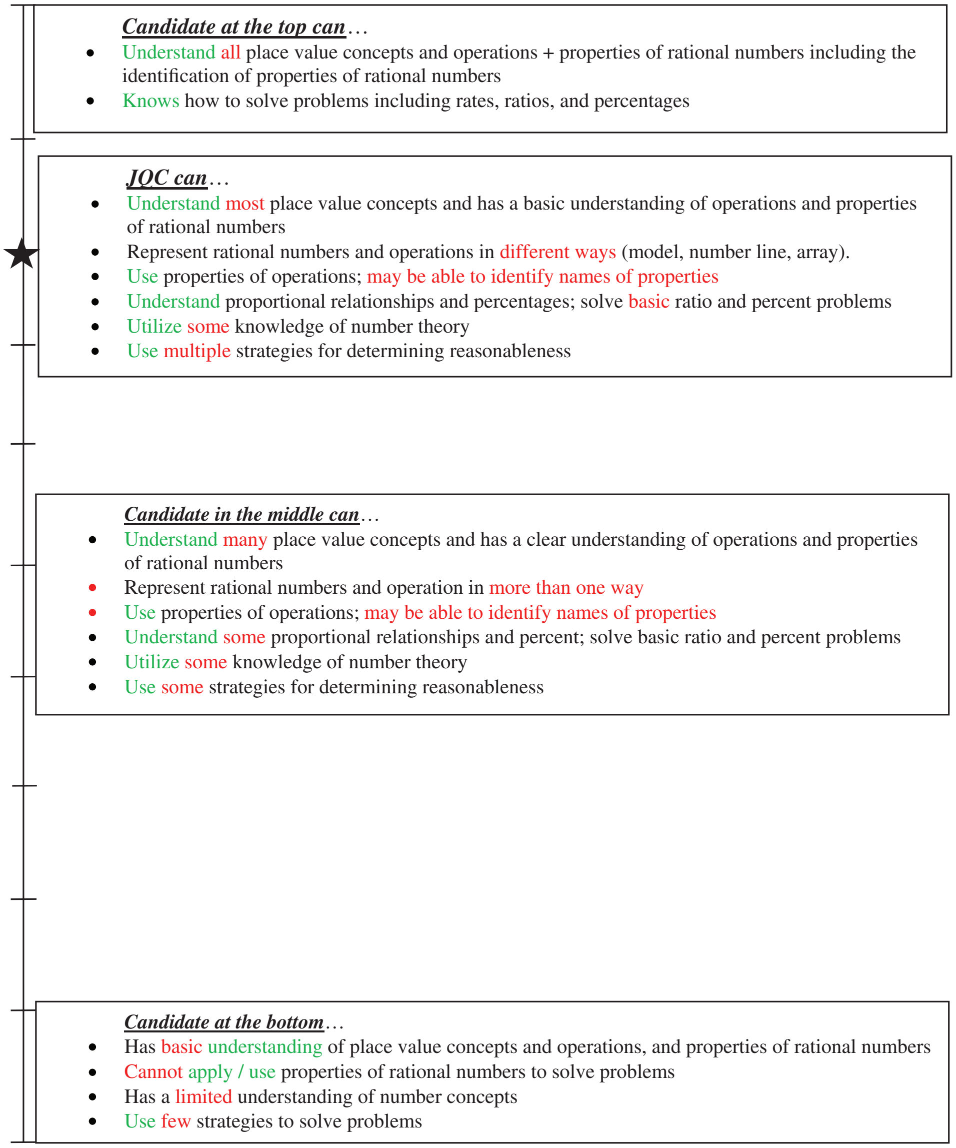

4
Persuades customers, managers, and coworkers about how to approach projects; asks detailed questions when assigned unclear tasks; carefully listens to clients,' managers,' and coworkers' opinions; works out disagreements with coworkers politely
3
Responds appropriately to coworkers' e-mails and voice mails; listens to criticism without becoming defensive; notices when clients and managers are unhappy about performance but does not directly address the issues; edits e-mails and memos for noticeable mistakes
2
Shares ideas with peers but occasionally without important details; usually listens to directions; notifies office when not coming to work for vacation and sick days; sometimes pays attention when clients or managers say they are upset
1
Responds aggressively when questioned; ignores dissatisfied customers; does not ask for clarification for unclear tasks; makes frequent, major errors when writing e-mails and documents that are sent to coworkers and clients

Figure 1 A hypothetical example of the behaviorally anchored rating scale for workplace communication skills.

to better contextualize the JQC and provide a frame of reference for the panelist in making standard-setting judgments. The fundamental principle of the AGR is to locate and then define the JQC, for each content domain, on a conceptual scale anchored by the expected test-relevant knowledge and/or skills of three benchmarks candidates: students with top, middle, and bottom performances (see Figure 2). These panelist-defined benchmarks serve as frames of reference against which the JQC is "slotted" and then are operationally defined by test-relevant knowledge and/or skills. We anticipate that explicitly organizing and situating the JQC in terms of a range of expected competence for benchmark candidates may offer a clearer target against which to make standard-setting judgments.

Furthermore, we make a deliberate effort to develop separate anchored representations for each major content domain measured by the test. In this regard, a profile of the JQC is established across the major test domains, potentially enabling the panelists to better understand the comparative strengths of the JQC by test domain (see Figure 3). For example, in one domain, it may be that the panelists located the JQC between the middle and top candidates (and defined accordingly), but for a second domain, the JQC is located between candidates with middle and bottom performances, as perhaps that second domain is, by its nature, a more challenging domain and to be considered just qualified as a beginning teacher, one would require comparatively less competence in that test domain.

We had conducted some preliminary research that supported the feasibility of constructing an AGR by standard-setting panelists and the ability of panelists to use it in making standard-setting judgments (Tannenbaum, Katz, & Kannan, 2015). In our previous study, 39 educators participated in three simulated standard-setting studies to compare the impact of text-based versus AGR-based JQC definitions. The educators were assigned to one of three conditions defined by the type of JQC descriptor to be created by the panel: text-create (created traditional list-based JQC), AGR-create (created AGR-based JQC), and AGR-receive (fleshed out the JQC portion of the AGR from the AGR-create group). Results from this study indicate that the text-based JQC resulted in higher cut-score recommendations than either AGR-based JQC. The group using the text-based JQC also displayed more lenient ratings when making inferences regarding the knowledge and skills a JQC may have beyond those explicitly included in the JQC description. However, the results of this study indicate that developing an AGR-based JQC versus receiving one and fleshing out the JQC portion did not seem to matter; the



Candidate at the top can…
- Understand all place value concepts and operations + properties of rational numbers including the identification of properties of rational numbers
- Knows how to solve problems including rates, ratios, and percentages

JQC can…
- Understand most place value concepts and has a basic understanding of operations and properties of rational numbers
- Represent rational numbers and operations in different ways (model, number line, array).
- Use properties of operations; may be able to identify names of properties
- Understand proportional relationships and percentages; solve basic ratio and percent problems
- Utilize some knowledge of number theory
- Use multiple strategies for determining reasonableness

Candidate in the middle can…
- Understand many place value concepts and has a clear understanding of operations and properties of rational numbers
- Represent rational numbers and operation in more than one way
- Use properties of operations; may be able to identify names of properties
- Understand some proportional relationships and percent; solve basic ratio and percent problems
- Utilize some knowledge of number theory
- Use some strategies for determining reasonableness

Candidate at the bottom…
- Has basic understanding of place value concepts and operations, and properties of rational numbers
- Cannot apply / use properties of rational numbers to solve problems
- Has a limited understanding of number concepts
- Use few strategies to solve problems

Figure 2 Example anchored graphical representation created by panelists for the numbers and operations domain in mathematics.

AGR-create and AGR-receive groups had similar cut-score recommendations and similar levels of inter-panelist consistency. The current study continues to explore the potential value of an AGR in comparison to a traditional (text-based) representation of the JQC for an educator licensure test—assessing readiness to be an elementary education teacher. Specifically, we evaluated if the pattern of passing scores (i.e., domain-level passing scores) differed among panelists who used different JQC definitions (i.e., AGR vs. text-based) and if the variance in their ratings differed. For the panels using the AGR, we expected that the passing scores would be more differentiated across domains and that the variance in ratings would be smaller, indicating that the AGR resulted in a differentiated (across domains), yet shared, understanding of the knowledge and skills of the JQC for the standard-setting panelists.

Method

In this paper, we describe results from two multi-panel standard-setting studies (i.e., for elementary education mathematics and elementary education social studies). In each study, we compared the impact of using different JQC definitions (i.e., AGR and traditional text-based descriptors) on panelists' standard-setting judgments. The elementary education mathematics and social studies tests from *THE PRAXIS SERIES*® assessments were identified as focal tests because both

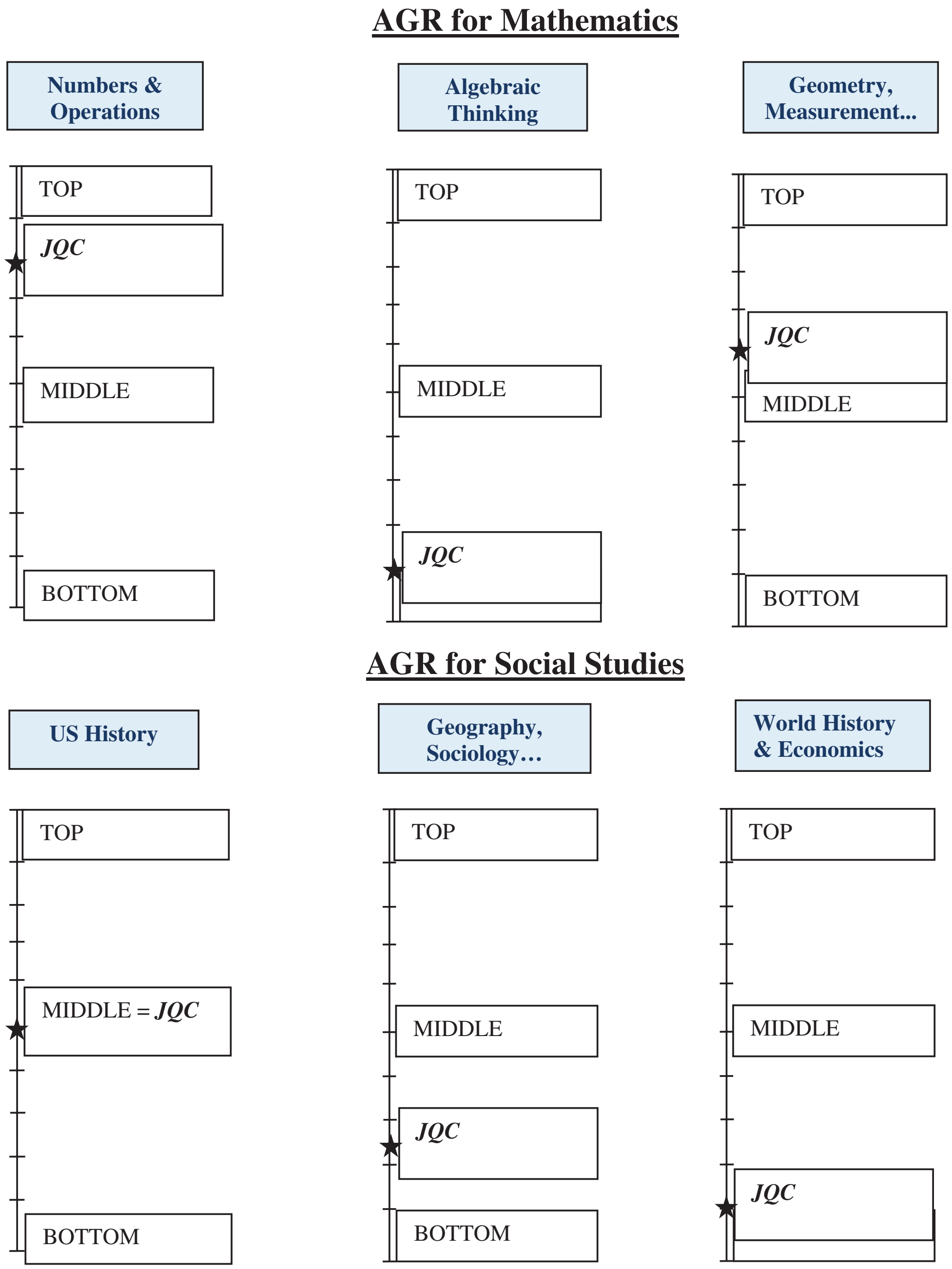


Figure 3 Panel-defined cross-domain anchored graphical representations (AGRs) for mathematics and social studies. Note. These AGRs represent the conceptual panel-defined locations for the borderline test taker across domains. The complete AGRs include descriptive ("can do") statements for each benchmark and borderline test taker level; see Figure 3 for an example AGR for one domain.

tests include three fairly differentiated content domains. Within each study (i.e., mathematics and social studies), two standard-setting panels were engaged in a mock (nonoperational) standard-setting study—the procedures for the two panels differed only in the type of JQC definition created and/or used.

Participants

Participants were certified elementary education mathematics ($n = 22$) and social studies ($n = 17$) teachers recruited from a database of elementary education teachers across New Jersey. In each study, participants were assigned to one of two conditions (i.e., panels), which were defined by the type of JQC definition used by the panel. In Study 1, 22 mathematics

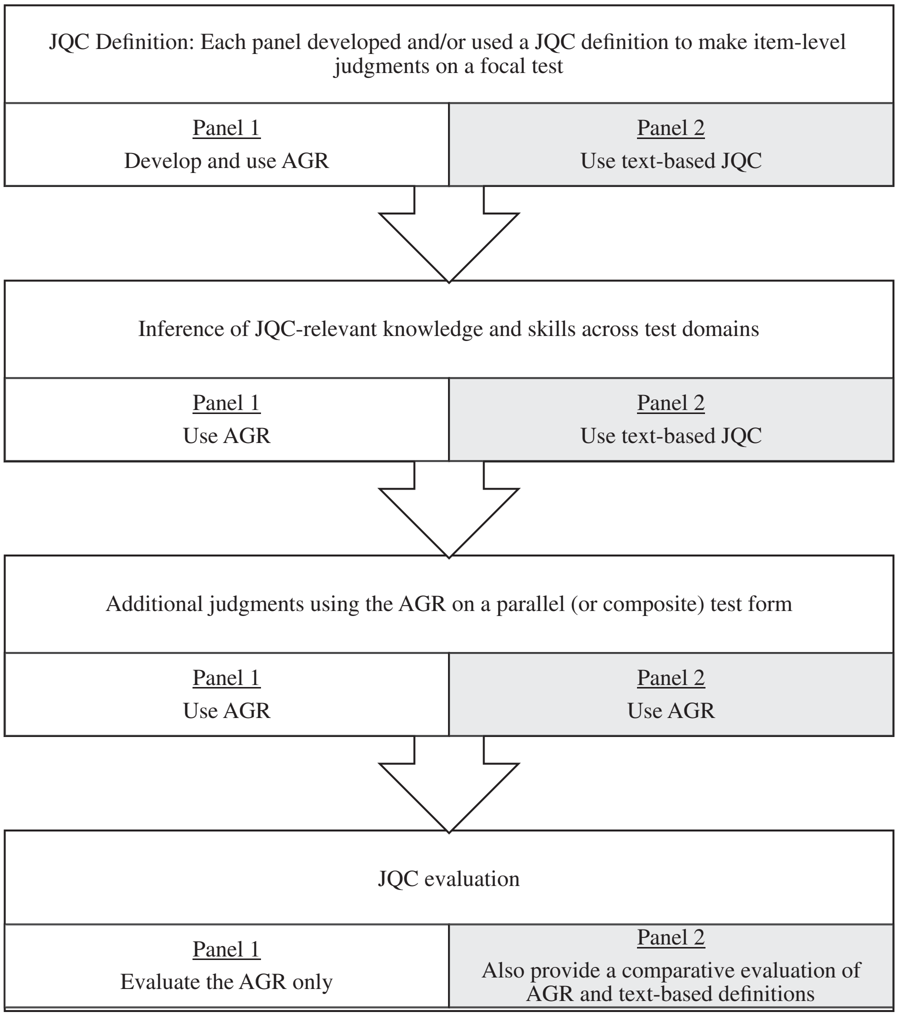

JQC Definition: Each panel developed and/or used a JQC definition to make item-level judgments on a focal test
Panel 1
Develop and use AGR
Panel 2
Use text-based JQC
Inference of JQC-relevant knowledge and skills across test domains
Panel 1
Use AGR
Panel 2
Use text-based JQC
Additional judgments using the AGR on a parallel (or composite) test form
Panel 1
Use AGR
Panel 2
Use AGR
JQC evaluation
Panel 1
Evaluate the AGR only
Panel 2
Also provide a comparative evaluation of AGR and text-based definitions

Figure 4 Study design for both studies. The same design was followed for both mathematics and social studies.

educators were assigned to either the AGR ($n = 11$) or traditional text-based ($n = 11$) condition. In Study 2, 17 social studies educators were assigned to either the AGR ($n = 8$) or text-based ($n = 9$) condition. Assignment was made randomly, with an attempt to distribute panelists evenly by gender, race, school district, and years of teaching experience. All participants received a $300 honorarium and a professional development certificate at the completion of the 2-day study.

Design and Procedures

Each study had two conditions: In the first (i.e., AGR) condition, the panel of educators created the JQC definition as three AGRs, one for each of the three test domains, and in the second condition, panelists used a traditional text-based JQC definition. The study design, along with all measures evaluated by panelists in both conditions, is presented in Figure 4.

Specifically, in the mathematics study, the first panel (i.e., the AGR panel) created three AGRs, one each for the domains (a) numbers and operations; (b) algebraic thinking; and (c) geometry and measurement, data, statistics, and probability. Panelists were divided into three subgroups. Each subgroup considered one domain of the test and created an AGR for that domain. Using the test content specification document, the panelists first defined the knowledge and skills of the anchor candidates, that is, candidates with top, middle, and bottom performances, in their assigned domain. They then slotted and defined the JQC in its appropriate location on the AGR. The whole panel reconvened to finalize all three AGR-based JQC definitions; panel-defined cross-domain locations of the JQC for mathematics are shown in Figure 3. Using this design, in which each subgroup developed the AGR for one test domain, which was then reviewed by the whole group, was both practical and robust, allowing us to develop AGRs for all test domains in 1 day. Figure 2 shows the detailed AGR for one domain, illustrated here for the numbers and operations domains in mathematics. It should be noted that, during training, panelists were instructed to use verb delimiters (e.g., understand vs. know) and modifiers (e.g., all vs. some) to clearly demarcate the knowledge and skills for the different levels (see Figure 2; delimiters are in green and modifiers are in red). This was intended to help them clearly understand what the JQC can and cannot do in comparison to the adjacent anchor candidates.

As indicated previously, the second panel used a text-based definition. To maintain consistency in the expectations of knowledge and skills for the JQC across the two panels, we extracted the bullet points defining the JQC level (from each domain of the AGR) and created a text-based (bulleted-list) definition that was given to the second panel. Through a whole-group discussion, this group refined the text-based JQC description to clarify or enhance it without altering its fundamental meaning (i.e., the knowledge and skill expectations for the JQC).

Similar procedures were followed in the second (social studies) study to create the JQC definition wherein the first (i.e., AGR) panel created three AGRs, one each for the domains (a) U.S. history, government, and citizenship; (b) geography, anthropology, and sociology; and (c) world history and economics. Panel-defined cross-domain locations of the JQC for social studies are also shown in Figure 3. Identical to the mathematics study, the second panel used a text-based JQC definition that was created from the AGR. In reviewing the definitions, Panel 2 (in both studies) made minimal changes to the language of the JQC.

Other than the JQC definition used, there were no procedural differences between the two panels in both studies. Both panels followed procedures typical to a probability-based Angoff standard-setting approach (Tannenbaum & Katz, 2013). After being introduced to the concept of standard setting, the panelists took the test; self-scored; and, as a group, discussed the knowledge and skills the test assessed. Once they had an understanding of the borderline test taker by either developing or refining a JQC definition, both panels received the same training on the probability-based Angoff approach and practiced making ratings. They made one round of judgments on the focal mathematics test (i.e., a 45-item test) in Study 1; however, in Study 2, panelists made two rounds of judgments (with intervening between-round discussions) on the focal social studies tests (i.e., a 55-item test).

Measures

In addition to evaluating the pattern of passing scores across the two study conditions, we wanted to evaluate if the AGR leads to more generalizable and consistent understanding of the borderline test taker for the panelists. In addition, we wanted panelists in each panel to evaluate the respective JQC definition they used to see if panelists self-report any procedural advantages in using the AGR when compared to a traditional text-based JQC definition. Specifically, both panels completed the following additional measures.

Additional Items

To investigate the panels' relative understanding of the JQC, they were asked to make standard-setting judgments on a set of novel test items. The use of unfamiliar items, rather than items that the panelists had previously reviewed, might "stress" panelists' understanding of the JQC. In addition, in both studies, Panel 2 was provided the AGR developed by the first panel after they had used the text-based description to make judgments on the focal test. Panel 2 used the AGR to make their standard-setting judgments on the additional items; this helped us evaluate if the introduction of the AGR (without having the ability to develop it) still allowed the second panel to internalize this definition — this would be reflected in the relative positioning of the JQC across domains in this panel's cut scores for the additional items.

We used a related PRAXIS elementary education licensure test (with 40 items) to make these judgments in mathematics. There were eight common items between Test 1 and Test 2. These items were used to evaluate consistency in panelist ratings across two occasions. For social studies, we created a composite form (with 80 items) from two additional forms of the elementary education social studies licensure test. We ensured that there were 25 common items between this composite form and the focal test, allowing us to evaluate panelist consistency in judgments across two occasions. Panel 1 (i.e., the panel that developed the AGR in both studies) continued to use the AGR definition in making their standard-setting judgments on these additional items.

Just Qualified Candidate Evaluation

We included multiple JQC evaluation measures at the end of the each study. Both panels (in each study) first completed a JQC evaluation measure for the respective JQC that was predominantly used by their panel. The purpose of this measure was to obtain feedback on the ease of use and clarity of each JQC definition. This measure was similar to a typical JQC evaluation measure used in any standard-setting study and included statements to indicate if the JQC was clear, easy to understand, and easy to use; the degree to which they had to make an inference; their confidence; and finally, their perceived accuracy of their ratings. The specific questions included in this measure are listed in Table 3 (along with results). Panel 2, in both studies, answered the last question (about perceived accuracy) after they had the opportunity to use the AGR. Next, because the text group (Panel 2 in both studies) had the opportunity to engage with both descriptors (i.e., AGR and text based), they were asked to complete a comparative evaluation of the two descriptors. The specific questions included in this measure are listed in Table 4 (along with results). Finally, because both panels had the opportunity to use the AGR by the end of each study, panelists from both panels provided some feedback on their perceived advantages and disadvantages of the AGR. Four open-ended questions were included in this measure; panelists were asked to describe their perceived advantages and disadvantages of the AGR, in addition to providing some feedback on how the AGR could be improved and if they thought there was a better way to represent the knowledge and skills of a JQC to help them with the standard-setting task.

Results

Results from the two multi-panel standard-setting studies, spanning two content areas (i.e., mathematics and social studies), are presented here. As discussed earlier, in each study, the first panel developed and then used an AGR to make their standard-setting judgments on the focal test, and the second panel was provided with a text-based bulleted list of the knowledge and skills for the JQC (extracted from the respective AGR developed by the first panel); both panels then used the AGR to make standard-setting judgments on a set of additional test items. In this section, we present the recommended cut scores (for overall test and by domain) and inter-panelist variability (i.e., standard deviations) for each panel in each study and present the consistency of their judgments on the additional test items. We also present results from the panels' self-reported evaluations of the respective JQC definition. Finally, because Panel 2 (i.e., the text-based group in each study) had the opportunity to engage with both JQC definitions, their comparative evaluation of the two JQCs are also presented here.

Recommended Cut Scores

Table 1 presents both the overall cut score and the cut score by test domain for each panel for the focal test form evaluated in each study. As a reminder, the two panels within each study used different JQC definitions for the focal test: Panel 1 used an AGR-based JQC definition, and Panel 2 used a text-based JQC definition. While the overall passing score did not noticeably vary between the conditions, consistent with our expectations, in both studies, the domain passing scores for the AGR groups reflected the relative positioning (see Figure 3) of the JQC for each test domain; the passing scores for the text groups were slightly less differentiated (Table 1). Across both studies, the inter-panelist variability (i.e., standard deviations) tended to be smaller for the AGR groups for each test domain (Table 1). However, it is interesting to note that the standard deviations for the social studies panels (both AGR and text) were somewhat larger than the standard deviations for mathematics, which may simply reflect the nature of the two content domains.

Consistency of Judgments on Additional Items

Educators from both groups completed standard-setting judgments on a set of additional test items of either an alternate form of 40 test items (mathematics) or a composite test form of 80 items created for this purpose (social studies). Panel 2 (in both studies) used the AGR to make their judgments on the additional items. These results are presented in Table 2. For both studies, the pattern of results in Table 2 shows that, once the AGR definition was provided, similar to Panel 1, Panel 2's cut scores also reflected the relative positioning of the JQC (see Figure 3) within each AGR domain. Moreover, in both studies, the standard deviations for Panel 2 (see Table 2) became smaller when compared to their inter-panelist

Table 1 Comparison of Passing Scores on the Focal Test Form

Panel (descriptor used)	Test domain	Proportional passing score	Interpanelist *SD*
Study 1 (elementary education mathematics test)			
Panel 1 (AGR)	*Overall test*	*.66*	*.17*
	Numbers and operations	.73	.16
	Algebraic thinking	.57	.18
	Geometry, etc.	.66	.16
Panel 2 (text)	*Overall test*	*.67*	*.19*
	Numbers and operations	.66	.20
	Algebraic thinking	.67	.19
	Geometry, etc.	.69	.17
Study 2 (elementary education social studies test)			
Panel 1 (AGR)	*Overall test*	*.57*	*.24*
	U.S. history	.61	.22
	Geography and sociology	.56	.25
	World history and economics	.53	.24
Panel 2 (text)	*Overall test*	*.57*	*.25*
	U.S. history	.50	.24
	Geography and sociology	.66	.26
	World history and economics	.61	.24

Note. See Figure 2 for the panel-defined conceptual locations of the borderline test taker for each test domain. AGR = anchored graphical representation.

Table 2 Comparison of Passing Scores on the Additional Test Form

Panel	Test domain	Proportional passing score	Inter-panelist *SD*
Study 1 (elementary education mathematics test)			
Panel 1	Numbers and operations	.73	.12
	Algebraic thinking	.64	.17
	Geometry, etc.	.70	.14
Panel 2	Numbers and operations	.71	.18
	Algebraic thinking	.62	.18
	Geometry, etc.	.68	.17
Study 2 (elementary education social studies test)			
Panel 1	U.S. history	.61	.21
	Geography and sociology	.59	.22
	World history and economics	.45	.23
Panel 2	U.S. history	.60	.23
	Geography and sociology	.58	.23
	World history and economics	.46	.24

Note. Both panels use anchored graphical representation. Educators in Panel 2 in each study had used the text-based descriptor when making their standard-setting ratings on the focal test.

variability on the focal test (presented in Table 1). However, it should be noted that the standard deviations for the social studies panels still remained larger than for the mathematics panels.

Consistency Measured Across Common Items

To evaluate if the AGR enables the panelists to be more consistent in their judgments, and to evaluate intra-panelist consistency, we compared the internal consistency with which each panelist made standard-setting judgments on the common items embedded in the two test forms. In the mathematics study, there were eight common items between Test 1 (i.e., the focal test) and Test 2 (i.e., the additional items). We used a nonparametric test owing to the small number of common items; we performed a repeated-measures Wilcoxon signed-rank test to determine the consistency in the rank ordering of these common items by each panelist. Results indicate that all panelists (across both panels) were internally consistent in their rank ordering of these eight common items. For the social studies study, we intentionally created a composite test

Table 3 Comparison of Ease of Use and Confidence in Making Judgments Using Both Just Qualified Candidates

Study	Panel 1 (AGR)	Panel 2 (text based)
Study 1 (elementary education mathematics test)		
1. The JQC definition was clear and easy to understand.	3.18 (0.41)	3.00 (0.45)
2. The JQC definition was easy to use.	3.00 (0.00)	3.00 (0.45)
3. The knowledge and skills measured by most of the items were explicitly covered in the JQC definition.	2.91 (0.54)	2.55 (0.52)
4. I had to make a huge inference about the JQC's performance on too many items.	2.00 (0.45)	2.55 (0.82)
5. I am confident about the item-level judgments I made using the JQC definition.	3.09 (0.30)	2.73 (0.47)
6. The JQC definition allowed me to make accurate judgments about the JQC's performance.	3.09 (0.30)	2.64 (0.51)
7. AGR only: Having the top, middle, and bottom anchors helped better orient my judgments.	3.64 (0.51)	3.64 (0.51)
Study 2 (elementary education social studies test)		
1. The JQC definition was clear and easy to understand.	3.00 (0.76)	3.33 (0.50)
2. The JQC definition was easy to use.	2.88 (0.64)	3.22 (0.44)
3. The knowledge and skills measured by most of the items were explicitly covered in the JQC definition.	2.50 (0.76)	2.67 (0.71)
4. I had to make a huge inference about the JQC's performance on too many items.	3.13 (0.35)	2.22 (0.44)
5. I am confident about the item-level judgments I made using the JQC definition.	3.00 (0.76)	3.11 (0.60)
6. The JQC definition allowed me to make accurate judgments about the JQC's performance.	3.00 (0.76)	3.22 (0.44)
7. AGR only: Having the top, middle, and bottom anchors helped better orient my judgments.	3.75 (0.46)	3.11 (0.60)

Note. Values are mean (*SD*). Higher values indicate more agreement. AGR = anchored graphical representation. JQC = just qualified candidate.

form with sufficient common items ($N = 25$) to be able to use a parametric method, such as correlation, to evaluate intra-panelist consistency with which each panelist made judgments on the 25 common items. On average, panelists in both panels (Panel 1, $r = .69$, $p < .05$; Panel 2, $r = .61$, $p < .05$) were internally consistent in their standard-setting judgments made on the common items. Therefore the type of JQC used did not appreciably impact internal consistency in panelist ratings across the two occasions.

Evaluation of the Alternate Just Qualified Candidate Definitions

At the end of each study, both panels (in both studies) completed multiple JQC evaluation forms. In the first of three measures, panelists provided feedback on the ease of use and comprehensibility of the specific JQC they used. These results are presented in Table 3; Panel 2 participants answered the last question after they had the opportunity to use the AGR. On average, mathematics educators in both panels agreed that their respective JQC definition was easy to understand and easy to use. Both mathematics panels somewhat disagreed with the statement that the knowledge and skills measured by the items were explicitly covered in the JQC definition. However, the two math panels significantly differed on their agreement to the following three statements: Significantly more Panel 2 (text-based) mathematics educators indicated that they had to make a huge inference about the JQC's performance on too many items, $F(1, 20) = 3.75$, $p < .05$. Moreover, significantly more mathematics educators in Panel 1 (AGR users) agreed that they were confident about their judgments, $F(1, 20) = 4.71$, $p < .05$, and that the JQC definition allowed them to make more accurate judgments, $F(1, 20) = 6.58$, $p < .05$.

However, the JQC evaluation results for social studies were starkly in contrast to the results observed for mathematics (results are presented in Table 3). On average, more social studies educators in Panel 2 than in Panel 1 thought that the JQC definition they used was clear, easy to understand, easy to use; indicated being more confident in their judgments; and perceived their judgments as more accurate. However, these differences were small. But it should be noted that significantly more social studies educators in Panel 1 (i.e., the AGR users) indicated that they had to make a huge inference about the JQC's performance on too many items, $F(1, 15) = 21.30$, $p < .001$, while significantly fewer educators in the text panel thought that having the top, middle, and bottom anchors helped to better orient their judgments, $F(1, 15) = 5.91$, $p < .05$.

Table 4 Comparative Evaluation of Both Borderline Test Taker Descriptors (Panel 2 Only)

Study	AGR	Text	Both equally
Study 1 (elementary education mathematics test)			
1. Which of the two descriptors helped you better understand the expected knowledge and skills	11	0	0
2. Which of the two descriptors did you find easier to use in making item-level judgments?	9	0	2
3. Which of the two descriptors helped you better understand if the knowledge/skills possessed by a borderline test taker met the demands of an item?	11	0	0
4. Which of the two descriptors required you to make less of an inference about a borderline test taker's performance on items?	11	0	0
5. Which of the two descriptors allowed you to be more confident about your judgments?	10	0	1
6. Which of the two descriptors allowed to you to make more accurate judgments?	10	0	1
7. Which of the two descriptors did you prefer?	11	0	0
Study 2 (elementary education social studies test)			
1. Which of the two descriptors helped you better understand the expected knowledge and skills	1	4	4
2. Which of the two descriptors did you find easier to use in making item-level judgments?	2	3	4
3. Which of the two descriptors helped you better understand if the knowledge/skills possessed by a borderline test taker met the demands of an item?	5	3	1
4. Which of the two descriptors required you to make less of an inference about a borderline test taker's performance on items?	6	3	0
5. Which of the two descriptors allowed you to be more confident about your judgments?	5	2	2
6. Which of the two descriptors allowed to you to make more accurate judgments?	5	2	2
7. Which of the two descriptors did you prefer?	6	3	0

Note. AGR = anchored graphical representation.

Because Panel 2 (in both studies) had the opportunity to engage with both JQC definitions, this panel was asked to complete a comparative evaluation of both JQCs. These results are presented in Table 4. It can be seen that across the board, mathematics educators indicated that the AGR enabled them to better understand the JQC's knowledge and skills, was easier to use, helped them to better determine if the JQC possessed the knowledge and skills measured by individual items, enabled them to make less of an inference, made them more confident about their judgments, enabled them to make more accurate item-level judgments, and helped them to better situate the JQC on the judgment scale. Finally, all 11 mathematics educators unequivocally indicated that they preferred the AGR over the text-based JQC definition. However, the results for social studies are not overwhelmingly in support of the AGR definition (see Table 4). Equal numbers of social studies educators thought either that the text-based definition helped them better understand the JQC and was easier to use or that the two definitions were equally helpful and easy. Though more social studies educators indicated that the AGR helped them to better determine if the JQC possessed the knowledge and skills measured by individual items, enabled them to make less of an inference, made them more confident about their judgments, enabled them to make more accurate item-level judgments, and helped them to better situate the JQC on the judgment scale, these preferences were not overwhelmingly in support of the AGR, as they were for the mathematics panel.

Finally, both panels (in both studies) were asked to respond to four open-ended questions on which they provided feedback about the advantages and disadvantages of the AGR. Table 5 summarizes some of the major advantages and disadvantages of the AGR panelists identified in each study. Most panelists in both studies considered the ability to compare the JQC with anchors, and having that frame of reference was one of the major advantages of the AGR. For example, one panelist said,

> The AGR clearly outlined the differences between the JQC and the candidates in the "top," "bottom," and "middle"—discussing these distinctions in our small groups as we developed the AGR, and then sharing those distinctions across the whole group, helped deepen my understanding of the JQC.

Moreover, panelists, particularly in Study 1, felt that the anchors/frames of reference helped them make more accurate judgments. For example, one panelist said that "having the AGR to refer to while making my judgments helped me make

Table 5 Summary of Advantages and Disadvantages of the Anchored Graphical Representation Identified by Panelists in Each Study

	No. of panelists	
Advantages/disadvantages	Study 1 (mathematics)	Study 2 (social studies)
Advantages		
Less inference/improved understanding and accuracy of judgments	9	2
Comparison with anchors/frame of reference	12	9
Visual representation helpful to understand differences between domains	6	3
Delimiters and modifiers help serve as a rubric	6	5
Disadvantages		
Too much reliance on position to make judgments	9	4
Teasing apart differences between JQC and adjacent anchors	6	2
Too much information to process/difficult to build	2	10
Too much inference	2	5

Note. Data from Panels 1 and 2 are combined and presented for the whole study. JQC = just qualified candidate.

better decisions," while another panelist said that "making judgments using the AGR made it a much more concrete, and less subjective decision."

Although approximately five panelists in each study indicated that they could not identify a significant disadvantage to the AGR, a majority of the panelists in Study 1 thought that one of the biggest disadvantages of the AGR was an overreliance on the cross-domain position to make their judgments. For example, one panelist said, "You tend to get stuck on using the same types of judgments for almost all items from one domain," and another panelist said, "Any algebraic thinking question automatically got a lower probability because JQC was placed lower on that domain." In contrast, in Study 2 (specifically for Panel 2 in the social studies study), the biggest identified disadvantage to the AGR was that there was too much information to process. A number of panelists (Panel 2, Study 2) indicated that they preferred the text-based definition because they felt overwhelmed with the amount of information on the AGR. For example, one panelist said, "Three separate pieces of paper to sort through (as opposed to a single sheet summarizing the knowledge and skills of the JQC in the text-based definition) can just be too overwhelming as you are making item-level judgments."

We also asked panelists how the AGR could be improved and if they thought there was a better way to represent the knowledge and skills of a JQC to help them with the standard-setting task. Though a majority of the panelists in Study 1 (i.e., the mathematics study) indicated that they thought the AGR was easy to use and could not think of a way to improve it, approximately five panelists particularly indicated that they would have liked more concrete examples in the AGR to specifically tease apart the knowledge and skills for the JQC and adjacent anchors. However, a number of panelists in Study 2 (i.e., the social studies study) thought that it was too cumbersome to build the AGR (Panel 1), that they would have liked the opportunity to revise the AGR between Round 1 and Round 2 judgments (Panel 1), and that adding more specific examples would have helped them better internalize the AGR (both Panel 1 and Panel 2).

Discussion

Standard-setting studies are conducted to identify cut scores that classify test takers into different levels of performance. In educator licensure testing, the goal is to differentiate between fail status and pass status. The minimum knowledge and skills required to reach or enter a level are operationally defined by a JQC description. Traditionally, a JQC includes a decontextualized list of knowledge and skills. In this study, we investigated an alternative, referred to as an AGR. An AGR is developed for each major content domain addressed by a test and includes a description of the knowledge and skills expected of students with top, middle, and bottom performances that serve as behavioral anchors in defining, understanding, and then applying the JQC description to make item-level judgments. We expected that this representation would provide a clearer frame of reference for making standard-setting judgments than the traditional representation and evaluated the AGR in two mock multi-panel standard-setting studies for mathematics and social studies educator licensure tests.

One of the challenges in standard-setting research is that there is no right or wrong cut-score recommendation; as Kane (2001) has noted, standard setting is about policies and policy formation, and as such, one needs to consider the reasonableness of processes and outcomes. It may be argued that an AGR provides a more reasonable description of a JQC by virtue of it being domain specific (and thereby reducing content underrepresentation) and its inclusion of behavioral anchors, which offer explicit “benchmarks” against which to consider the JQC. If it is reasonable to expect that a passing score in licensure may be obtained by demonstrating more competence on certain subdomains than others, then explicitly organizing the JQC description by content domain and in terms of contextualized expectations for competence in each domain may offer a clearer frame of reference against which to make standard-setting judgments.

Our collective results from two studies reinforce that there are differences between standard-setting judgments made using a traditional JQC description and those made using the AGR. Even though the overall passing scores did not differ between the panels, we consistently found that using the AGR resulted in standard-setting ratings (at the test domain level) that are more aligned with the panelists’ differential expectations of the JQC’s performance for each domain. This suggests that using the text-based JQC definition limits the extent to which raters are thinking about each item in the context of their overall expectation for a JQC in that specific subarea, while explicitly situating and then defining the JQC in the context of the anchors in an AGR helps with the internalization of those differential expectations, which is reflected in their judgments. Additional validity evidence for the reasonableness of the anchors in supporting panelist judgments may be found in the panelists’ open-ended feedback about the advantages of using the AGR. Panelist feedback across both studies overwhelmingly indicated that the anchors provided them with the necessary frame of reference in making their standard-setting judgments, and panelists (particularly in Study 1) perceived that the AGR enabled them to be more confident in their judgments and helped them make more accurate judgments.

Limitations and Practical Implications

A well-constructed JQC descriptor is needed to arrive at a reasonable passing score (Egan et al., 2012; Perie, 2008). The collective results from our studies indicate that visually organizing the AGR by subdomain and including performance anchors for benchmark candidates resulted in standard-setting judgments that were not only more consistent with panelists’ differential expectations across domains but also resulted in slightly smaller variability across panelists and an increased perception of confidence and accuracy in their own judgments. However, there are some limitations to the generalizability of the findings from this study and practical feasibility of the proposed new method; these limitations are noted here.

First, within the constrained design of this study, it should be noted that the text-based definition of the JQC was derived from the AGR. Therefore the two definitions are not truly independent of each other, and the JQC knowledge and skill expectations of the text-based group were, in part, influenced by the AGR group. Moreover, the text-based group was also asked to consider the AGR to rate novel items after having used a text-based definition first. Even though the text-based definition was derived directly from the AGR, we should acknowledge that using two different types of definitions might possibly have confounded the results for Panel 2. If the text-based group had independently developed their own JQC definition, then it is possible that their expectations of the JQC knowledge and skills may have been very different, resulting in much greater differences between the groups in their overall cut scores and in the resultant differential expectations across the domains. Follow-up studies should evaluate the differences in resultant cut scores when the two definitions are developed independently. In addition, it is recommended that follow-up studies also evaluate the recommended cut scores for each group using the item response theory theta metric for each domain and also evaluate the correlation between the item ratings and conditional p-values, if feasible, to further substantiate the results.

Second, the task of developing the AGR by subdomain and internalizing this description might be time consuming and tedious when undertaken within the context of a typical standard-setting meeting, in which panelists also have the rather arduous task of making probability judgments at the item level across multiple rounds. Several panelists, particularly in Study 2, thought that the AGR was difficult (and, perhaps, time consuming) to build. However, results from both studies show that Panel 2 was able to apply the AGR as successfully (as evident from the pattern of domain cut scores on the additional judgments) as the panel that developed the AGR. These results indicate that standard-setting panelists are able to internalize and apply the AGR even if they were not involved in developing the AGR-based JQC description. In our own previous evaluations of the AGR (Tannenbaum et al., 2015), we have found that cut score estimates did not differ significantly for a panel that developed the AGR and another panel that applied the AGR.

As a practical feasibility issue, it is possible that the AGRs across domains were not equally fleshed out across domains because the task of developing the AGRs was condensed into an abbreviated activity on the first day during a standard-setting meeting. In fact, several panelists (across both studies) indicated that the AGRs for specific domains, developed in small groups, were not consistent in quality or amount of information included. For example, the AGR for one subdomain (e.g., numbers and operations) may not have been as elaborate and clearly crafted as the AGR for another subdomain (e.g., algebraic thinking), and panelists found that this inconsistency hindered them from applying the AGR equally well across test domains. Therefore, one possible solution would perhaps be to build the AGR in a separate meeting (such as the PLD meetings conducted within K–12), which could then be internalized by the standard-setting panelists and applied to make item-level judgments. If the AGR is developed in a separate meeting, the amount of information and content of the AGR may be standardized across domains before providing it to the standard-setting panelists, who may then be able to appropriately internalize, understand, and apply the description equally well for all subdomains on the test.

Finally, it should be noted that these results are based only on two small-scale, mock standard-setting meetings conducted for two elementary education licensure tests. Moreover, in one of the two studies (i.e., social studies), panelists found that the AGR was more cumbersome to build and use. It is possible that the process of building and using the AGR is too complicated for some content domains so that the potential benefit is overweighed by the cognitive demand of building and using a complex JQC description. However, this could also be a panel effect, because we have only evaluated the AGR for two content domains with two panels each. Therefore, the value of developing and employing a graphical JQC description like the AGR must be further evaluated for other licensure assessments.

Conclusions

In standard-setting, the integrity of the means is critical to evaluating the reasonableness of the ends—the passing score recommendation (Kane, 2001; Tannenbaum & Katz, 2013). Collectively, though we found that the test-level passing scores were comparable, our results indicate that use of the AGR resulted in standard-setting ratings (at the test domain level) that were more aligned with the panelists' differential expectations of the JQC's performance for each domain. We believe that this alignment adds to the procedural validity evidence supporting the reasonableness of standard-setting recommendations. However, the qualitative feedback from the social studies panel suggests that some panelists might have preferred the text-based definition. Therefore, additional evaluation of the value of using the AGR for various subject areas should be considered. Though it is the case that in any standard setting, it is not possible to evaluate the accuracy of the cut scores, but rather it is only possible to evaluate the reasonableness of the cut score development process, nevertheless, the cut scores derived from these two methods should be evaluated for consistency with student performance data to determine if one method leads to cut scores that are more clearly aligned with actual student performance. Finally, it should be noted that our results are based on two small-scale mock implementations of standard setting for licensure tests; though promising, results from our studies should be generalized with caution.

References

Bauer, M. I., & Johnson-Laird, P. (1993). How diagrams can improve reasoning. *Psychological Science, 4*, 372–378. https://doi.org/10.1111/j.1467-9280.1993.tb00584.x

Bernardin, H. J., & Beatty, R. W. (1984). *Performance appraisal: Assessing human behavior at work*. Boston, MA: Kent.

Bernardin, H. J., & Buckley, M. R. (1981). A consideration of strategies in rater training. *Academy of Management Review, 6*, 205–212.

Campbell, J. P., Dunnette, M. D., Arvey, R. D., & Hellervick, L. V. (1973). The development and evaluation of behaviorally anchored rating scales. *Journal of Applied Psychology, 57*, 15–22. https://doi.org/10.1037/h0034185

Chung, H. M., & Khan, M. B. (2008). Classification of unethical behaviors in the management of information systems: The use of behaviorally anchored rating scale procedures. *International Journal of Management, 25*, 262–269.

Cizek, G. J. (1993). Reconsidering standards and criteria. *Journal of Educational Measurement, 30*, 93–106. https://doi.org/10.1111/j.1745-3984.1993.tb01068.x

Cizek, G. J. (2012). An introduction to contemporary standard setting: Concepts, characteristics, and contexts. In G. J. Cizek (Ed.), *Setting performance standards: Foundations, methods, and innovations* (2nd ed., pp. 3–14). New York, NY: Routledge.

Cizek, G. J., & Bunch, M. B. (2007). *Standard setting: A guide to establishing and evaluating performance standards on tests*. Thousand Oaks, CA: Sage. https://doi.org/10.4135/9781412985918

Cook, S. S. (1989). Improving the quality of student ratings of instruction: A look at two strategies. *Research in Higher Education, 30*, 31–45. https://doi.org/10.1007/BF00992789

Day, D. V., & Sulsky, L. M. (1995). Effects of frame-of-reference training and information configuration on memory organization and rating accuracy. *Journal of Applied Psychology, 80*(1), 158–167. https://doi.org/10.1037/0021-9010.80.1.158

Diekhoff, G. M., Brown, P. J., & Dansereau, D. F. (1982). A prose learning strategy training program based on network and depth-of-processing models. *Journal of Experimental Education, 50*, 180–184. https://doi.org/10.1080/00220973.1982.11011820

Egan, K. L., Schneider, M. C., & Ferrara, S. (2012). Performance level descriptors: History, practice, and a proposed framework. In G. J. Cizek (Ed.), *Setting performance standards: Foundations, methods, and innovations* (2nd ed., pp. 79–106). New York, NY: Routledge.

Gobert, J. D., & Clement, J. J. (1999). Effects of student-generated diagrams versus student-generated summaries on conceptual understanding of casual and dynamic knowledge in plate tectonics. *Journal of Research in Science Teaching, 36*, 39–53. https://doi.org/10.1002/(SICI)1098-2736(199901)36:1<39::AID-TEA4>3.0.CO;2-I

Hall, R. H., Hall, M. A., & Saling, C. B. (1999). The effects of graphical postorganization strategies on learning from knowledge maps. *Journal of Experimental Education, 67*, 101–112. https://doi.org/10.1080/00220979909598347

Hambleton, R. K. (2001). Setting performance standards on educational assessments and criteria for evaluating the process. In G. J. Cizek (Ed.), *Setting performance standards: Concepts, methods, and perspectives* (pp. 89–116). Mahwah, NJ: Lawrence Erlbaum.

Hambleton, R. K., & Pitoniak, M. J. (2006). Setting performance standards. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 433–470). Westport, CT: American Council on Education/Praeger.

Impara, J. C., Giraud, G., & Plake, B. (2000, April). *The influence of providing target group descriptors when setting a passing score*. Paper presented at the meeting of the American Educational Research Association, New Orleans, LA.

Kane, M. T. (2001). So much remains the same: Conception and status of validation in setting standards. In G. J. Cizek (Ed.), *Setting performance standards: Concepts, methods, and perspectives* (pp. 53–88). Mahwah, NJ: Erlbaum.

Perie, M. (2008). A guide to understanding and developing performance-level descriptors. *Educational Measurement: Issues and Practice, 27*, 15–29. https://doi.org/10.1111/j.1745-3992.2008.00135.x

Plake, B. S., Impara, J. C., & Irwin, P. (1999). *Validation of Angoff-based predictions of item performance*. Retrieved from ERIC database. (ED430004)

Pounder, J. S. (2000). A behaviorally anchored rating scales approach to institutional self-assessment in higher education. *Assessment & Evaluation in Higher Education, 25*, 171–182. https://doi.org/10.1080/713611422

Pulakos, E. D. (1984). A comparison of rater training programs: Error training and accuracy training. *Journal of Applied Psychology, 69*, 581–588. https://doi.org/10.1037/0021-9010.69.4.581

Schleicher, D. J., Day, D. V., Mayes, B. T., & Riggio, R. E. (2002). A new frame for frame-of-reference training: Enhancing the construct validity of assessment centers. *Journal of Applied Psychology, 87*, 735–746. https://doi.org/10.1037/0021-9010.87.4.735

Schwab, D. P., Heneman, H. G., & DeCotiis, T. A. (1975). Behaviorally anchored rating scales: A review of the literature. *Personnel Psychology, 28*, 549–562. https://doi.org/10.1111/j.1744-6570.1975.tb01392.x

Smith, P. C., & Kendall, L. M. (1963). Retranslation of expectations: An approach to the construction of unambiguous anchors for rating scales. *Journal of Applied Psychology, 47*, 149–155. https://doi.org/10.1037/h0047060

Stoskopf, C. H., Glik, D. C., Baker, S. L., Ciesla, J. R., & Cover, C. M. (1992). The reliability and construct validity of a behaviorally anchored rating scale used to measure nursing assistant performance. *Evaluation Review, 16*, 333–345. https://doi.org/10.1177/0193841X9201600307

Stull, A., & Mayer, R. E. (2007). Learning by doing versus learning by viewing: Three experimental comparisons of learner-generated versus author-provided graphic organizers. *Journal of Educational Psychology, 99*, 808–820. https://doi.org/10.1037/0022-0663.99.4.808

Sulsky, L. M., & Kline, T. J. B. (2007). Understanding frame-of-reference training success: A social learning theory. *International Journal of Training and Development, 11*, 121–131. https://doi.org/10.1111/j.1468-2419.2007.00273.x

Tannenbaum, R. J. (2011). *Setting standards on THE PRAXIS SERIES™ tests: A multistate approach* (R&D Connections No. 17). Princeton, NJ: Educational Testing Service.

Tannenbaum, R. J., & Kannan, P. (2015). Consistency of Angoff-based standard-setting judgments: Are item judgments and passing scores replicable across different panels of experts? *Educational Assessment, 20*, 66–78. https://doi.org/10.1080/10627197.2015.997619

Tannenbaum, R. J., & Katz, I. R. (2013). Standard setting. In K. F. Geisinger (Ed.), *APA handbook of testing and assessment in psychology: Vol 3. Testing and assessment in school psychology and education* (pp. 455–477). Washington, DC: American Psychological Association. https://doi.org/10.1037/14049-022

Tannenbaum, R. J., Katz, I. R., & Kannan, P (2015). *Anchored graphical representations: An alternative to traditional performance level descriptors*. Paper presented at the meeting of the National Council on Measurement in Education, Chicago, IL.

Uggerslev, K. L., & Sulsky, L. M. (2008). Using frame-of-reference training to understand the implications of rater idiosyncrasy for rating accuracy. *Journal of Applied Psychology, 93*, 711–719. https://doi.org/10.1037/0021-9010.93.3.711

Woehr, D. J., & Huffcutt, A. I. (1994). Rater training for performance appraisal: A quantitative review. *Journal of Occupational and Organizational Psychology, 67*, 189–205. https://doi.org/10.1111/j.2044-8325.1994.tb00562.x

Suggested citation:

Kannan, P., Tannenbaum, R. J., & Hebert, D. (2018). *Anchored graphical representations: A graphical alternative to traditional just qualified candidate descriptors for licensure tests* (Research Report No. RR-18-40). Princeton, NJ: Educational Testing Service. https://doi.org/10.1002/ets2.12228

Action Editor: James Carlson

Reviewers: Spiros Papageorgiou and Caroline Wiley

ETS, the ETS logo, MEASURING THE POWER OF LEARNING, and THE PRAXIS SERIES are registered trademarks of Educational Testing Service (ETS). All other trademarks are property of their respective owners.

Find other ETS-published reports by searching the ETS ReSEARCHER database at http://search.ets.org/researcher/