Educational Measurement
ISSUES AND PRACTICE

NCME national council on measurement in education

Educational Measurement: Issues and Practice
Spring 2015, Vol. 34, No. 1, pp. 22–33

Gauging Item Alignment Through Online Systems While Controlling for Rater Effects

Daniel Anderson, Shawn Irvin, Julie Alonzo, and Gerald A. Tindal, *University of Oregon*

The alignment of test items to content standards is critical to the validity of decisions made from standards-based tests. Generally, alignment is determined based on judgments made by a panel of content experts with either ratings averaged or via a consensus reached through discussion. When the pool of items to be reviewed is large, or the content-matter experts are broadly distributed geographically, panel methods present significant challenges. This article illustrates the use of an online methodology for gauging item alignment that does not require that raters convene in person, reduces the overall cost of the study, increases time flexibility, and offers an efficient means for reviewing large item banks. Latent trait methods are applied to the data to control for between-rater severity, evaluate intrarater consistency, and provide item-level diagnostic statistics. Use of this methodology is illustrated with a large pool (1,345) of interim-formative mathematics test items. Implications for the field and limitations of this approach are discussed.

Keywords: interim-formative assessment, item-standard alignment, rater effects, Rasch modeling, test Development, large-scale testing,

Item alignment to standards is crucial for establishing the validity of standards-based test inferences (La Marca, 2001), particularly given that test-based accountability policies, such as the *No Child Left Behind Act* (NCLB, 2001), depend upon the alignment of items to standards in drawing inferences on the effectiveness of educational institutions. Alignment to standards is also important for formative evaluation systems, because the validity of teachers' instructional decisions depends, in part, upon the degree of item alignment. Misaligned items may lead to misguided instructional decisions. In the end, standards can then be used to align instruction with accountability.

During typical alignment studies, multiple raters judge individual test items on various dimensions. The process either entails reviews that are completed mostly independently (Porter, 2002, 2004) or by panels of experts convening for collaborative, discussion-based reviews (Rothman, Slattery, Vranek, & Resnick, 2002; Webb, 1999, 2007). With each approach, the ratings from multiple reviewers are eventually combined. Porter and Webb recommend averaging the ratings, while Rothman et al. (2002) recommend the panel of experts continue discussion until a single *consensus* rating is established.

In this paper, we focus on the process of gauging item-standards alignment through online methods that do not require raters to convene in person. The proposed methodology provides flexibility for test developers to gather experts from a broad geographical base, reduces costs so that more raters can potentially be assigned to reviews, and allows for large item banks to be more efficiently reviewed by extending the review time to weeks rather than days. Ratings from multiple reviewers are then combined through latent trait methods that statistically control for rater severity and can be used to evaluate intrarater reliability, rather than computing averages or gaining consensus. We illustrate our method with data from a large pool (1,345) of interim-formative mathematics assessment items. The panel method is first reviewed and then contrasted with online designs. We then consider the dynamics of discussion-based reviews, and why it may be desirable to avoid these dynamics when possible.

Panel Alignment Methods

When panel methods are used, groups of content-matter experts (e.g., classroom teachers or curriculum specialists) are convened to judge the alignment of test items to standards (Rothman et al., 2002). In his influential report, Webb (1999) describes a panel methodology to establish the alignment of test items with content standards that is completed through "a 4-day Alignment Analysis Institute" (p. 2), where raters judged the alignment of statewide math and science assessments from two or three grade levels. He also notes that the majority of raters attended a 1-day meeting prior to the institute, where they were introduced to the alignment process. In later reports, Webb (2007) notes that a 3-day process is perhaps more typical, but that the length of the institute depends on "the number of grades to be analyzed, the length of the standards, the length of the assessments, and the number of assessment forms under consideration" (pp. 8–9). He recommends 5–8 raters for each panel, noting that an increase in raters generally increases the reliability of ratings.

During the alignment process outlined by Webb (1999, 2007), raters first independently rate the degree of

Daniel Anderson 175 Lokey Education 5262 University of Oregon, Eugene, OR 97403-5262; daniela@uoregon.edu.

item-standard alignment and then discuss their ratings with others on the panel to form a deeper understanding of standards, items, and their relation. Following the discussion, raters are provided the opportunity to modify their ratings. Finally, ratings from the multiple reviewers are averaged to document whether predetermined alignment criteria are met, with standard deviations reported as an indicator of the variance between raters.

Rothman et al. (2002) proposed the *Achieve* model, which shares some procedural similarities with Webb's model: Panels of experts are convened and work first individually and then collaboratively in developing understanding around the alignment of items and standards. However, while the ratings from the multiple reviewers are averaged under Webb's model (1999, 2007), the *Achieve* model stipulates that a single consensus rating must be reached through group discussion (Case, Jorgensen, & Zucker, 2004; Martone & Sireci, 2009). Rothman et al. note that the process can be quite expensive and that the pool of raters should "represent a diversity of viewpoints" (p. 8), to ensure alignment ratings are not biased toward certain groups.

Online Versus Panel Methods

Although the panel method has been well documented (Council of Chief State School Officers [CCSSO], 2002), it also may be impractical in a number of situations, particularly if the pool of items to be reviewed is large, or if the individuals needed for their expertise are broadly distributed geographically. In interim-formative or computer-adaptive assessment systems the total item bank may contain several thousand items within each grade, making it even more difficult to convene raters in person. The financial and practical costs of bringing expert reviewers together to judge the alignment of even a portion of these item banks becomes, in most cases, unmanageably large. Furthermore, in many less densely populated states or regions, it may be difficult to bring together a sufficient number of content experts to form panels that encompass a diversity of viewpoints without considerable funds allocated to travel.

When judging the alignment of test items, it is critical that the participating raters represent the diversity of the intended users. "National" tests are increasingly common, and are likely to continue to increase in production as common standards are adopted (i.e., the Common Core State Standards [CCSS]). Obtaining participants from a wide-range of contexts increases the diversity of views (Rothman et al., 2002), reduces the chances that alignment ratings contain contextual biases, and increases the external validity of results. Tests designed for use across collaborative states necessitate alignment designs that sample raters from across the full user base; with high-stakes tests perhaps needing stratified random sampling techniques to ensure adequate representation.

The financial cost of such designs can be considerable. For example, consider a scenario in which roughly 3,000 items are in need of alignment review, such as in a computer adaptive testing context, in which a nationally representative sample of raters is needed (i.e., the test has a multistate user base). With common assessments across states becoming more regular, including many for high-stakes purposes (e.g., Smarter Balanced Assessment Consortium, 2012), the necessity of such designs is increasingly common. Webb (1999) included 16 raters to judge the alignment of a total of 859 items over the course of 4 days. If we assume that on average 1,000 items could be reviewed by 16 raters over 4 days, then the length of the institute would need to be extended to 12 days or the number of raters would need to be multiplied by three (48 raters) to keep the institute to 4 days. Balancing these extremes, we could recruit 24 raters for a 6-day institute, and estimate the cost of the design by assuming approximately half the reviewers would need flights (at ~$500) and nearly all would need a hotel room (~$150 per night) and food (~$75 per day). If we compensated raters at $150 per day, the cost of the study would be roughly $60,000, not including any facility costs. Conference rooms suitable for 24 people can be rented for approximately $500 per day, bringing the total cost of the hypothetical panel design to roughly $63,000.

Online designs make alignment reviews of large item banks more financially feasible by eliminating the majority, if not all costs relative to travel, lodging, and food. The variable costs (travel, lodging, and food) are much greater than the fixed costs (raters). For example, in the above scenario, we could easily include 24 reviewers and extend the time allowed for the review, reducing the potential confound of reviewer exhaustion, because reviewers would be able to review items independently at times that work best for them. Under this design, the total cost of the study would be roughly $21,600 (about one-third the estimated cost of the previously described design), assuming raters were paid ~$150 per day for the equivalent of 6 full days of work.

Online approaches can incorporate collaborative/discussion-based reviews, similar to in-person designs (Webb, Alt, Ely, & Vesperman, 2005), or reviews that are mostly or fully independent, where each rater provides judgments free of all other raters (Behavioral Research and Teaching, 2013; Porter, 2004). The data collection process through online designs need not differ from data collected from in-person designs, but the overall process becomes more financially feasible and logistically flexible.

Group Dynamics in Panel Methods

The panel method is thought to enhance interrater reliability through a process of raters defending their initial judgments (elaborating on their thought processes) and then potentially modifying their ratings after group discussion to reach a final judgment (see Rothman et al., 2002; Webb, 1999). Raters are thought to gain a deeper understanding of items and standards through sharing information, which raters may use to modify their initial judgments. Group discussion also facilitates convergence: Raters who are overly severe or lenient are more likely to be reined into conformity by the panel and subsequently (and hopefully appropriately) modify their ratings (Kocher & Sutter, 2005). By design, however, all panel models provide opportunities for a single rater to affect the judgment of others. An unintended consequence of this design is that an outlying rater may sway others in a manner that inappropriately over- or underrepresents the alignment of the item.

Persuasion is an essential component of within-group dynamics present in panel settings (Brown, 1988), and an overbearing person may have substantial influence over the final item rating, while the input of a more unobtrusive rater may have less. The process of persuasion and subsequent conformity occurs regularly in such settings. In a seminal

study on conformity in groups, Asch (1956) found that the rate of conformity in small groups rose rapidly when individuals faced a group of two or three. More recently, in a meta-analysis of studies replicating Asch's experiments, Bond (2005) found largely the same pattern of small group conformity. If an individual in a small group setting convinces one or two individuals to move in his or her direction, the opinion of the group may follow, with conformity/consensus the end product. If such an outcome were to occur in an alignment process, this may be an advantage if the influential rater is pulling others in a direction that more appropriately represents the alignment of an item. However, the influential rater may just as easily pull the group in a manner that under- or overrepresents the alignment of the item.

Online interactions likely reduce the rate of conformity when compared to in-person reviews, because many of the situational and social pressures are removed. For instance, Laporte, van Nimwegen, and Uyttendaele (2010) replicated Asch's (1956) experiment but within an online environment. The authors found that the rate of conformity was lower when individuals only saw a picture of the other "participants" (all of whom were confederates in the study) than when all participants could see each other through a video monitor. Across conditions, however, the overall level of conformity was lower than is typically found. These findings would seem to suggest that even with online models where a discussion component exists (e.g., Webb et al., 2005), the overall rate of conformity would likely be reduced because many of the situational and social pressures are removed.

Online Reviews and Possibilities for Scaling

In the alignment methodology we describe, online reviews are conducted with raters providing judgments independently (i.e., without group discussion). We opted for independent ratings as opposed to discussion-based ratings to eliminate the potential for group dynamics negatively influencing the final ratings, while also recognizing that any potential benefits are also removed. For example, independent reviews do not provide opportunities for extreme ratings to be tempered, one of the intended purposes of discussions in panel designs. Overly severe or lenient raters can have substantial practical repercussions on an alignment study (Wolfe, 2004). False negative items (those mistakenly rated as not aligned) can waste valuable test-developer resources, while false positive items (items mistakenly rated as aligned) threaten the validity of standards-based test inferences.

To help document rater consistency and control for rater severity, latent trait methods can be applied to alignment data. Test developers can then capitalize on the benefits of an online design (i.e., cost, efficiency, flexibility), while better evaluating the internal consistency of raters and correcting for raters' lenient or severe tendencies that threaten validity. The many-facets Rasch model (MFRM) is a commonly employed statistical model for accounting for rater severity (e.g., Engelhard, 1994; Myford & Wolfe, 2000). Under the MFRM, raters are treated as a facet of the measurement process with *conditions* of measurement equated—provided raters judge common items. By equating the conditions of measurement, all items are placed on a common scale, taking into account the specific rater who judged each item. Ratings provided by overly severe raters are adjusted toward the *aligned* end of the scale, while ratings provided by overly lenient raters are adjusted toward the *unaligned* end of the scale. Concerns of interrater reliability are then mitigated, as the analysis statistically corrects for differences between raters in terms of severity; and intrarater reliability becomes the primary concern.

As we discuss in the Methods section, the MFRM must be slightly redefined to accommodate alignment data, where *items*, rather than persons, become the object of measurement. Rather than items indicating the location on the latent trait for persons, raters are the indicator of the latent trait for items. Because conditions of measurement are equated, item locations are conditioned on the severity of each reviewer who rates the item. The internal consistency of each rater is evaluated by examining the residual between a model-based expected value and an observed value. The expected value is calculated based on (1) how the rater judged all other items (i.e., the severity of the rater), and (2) how all other raters judged the item in question (i.e., the "difficulty" of endorsing the item as aligned). If the expected value closely matches the observed value, then the rater is operating consistently. A similar indicator of rater agreement for each item also can be calculated.

In this study, we apply a latent trait model to data collected from an online alignment review. The study is intended to illustrate a methodology that provides test developers an efficient means to gauge the alignment of large item banks, gather raters from diverse geographical regions, and reduce the overall cost of alignment studies. The specific characteristics of our alignment review follow.

Methods

The context of our application of latent trait modeling was a study of the alignment between interim-formative assessment mathematics item content and the CCSS. We primarily address the manner in which we selected items and standards, trained raters, and developed our rating scale; we end the methods section with a complete description of the analytic procedures.

Participant Recruitment and Item-Standard Selection

Fifteen middle school math teachers from across the United States were recruited to serve as content expert raters. All raters had experience with the CCSS and experience instructing middle school students in mathematics, 10 held an advanced degree, nine held a mathematics specialization and/or worked as a district math coach, and six primarily instructed students receiving special education services.

The full item bank used in this study included 900 items in each of Grades 6–8, totaling 2,700 unique items intended for use in a national interim-formative assessment system. To help sample adequately across all item types in the full item bank, items were selected for review based on a randomized matrix-sampling plan, stratified by the original item-writer and the CCSS to which the item was written. A total of 1,345 items were selected for review (approximately 50% of the total item bank), with each rater asked to judge the alignment of approximately 270 items.

All items were written to align with only one standard during item development. In an early step in the alignment study process, the targeted CCSS of each item was uploaded to an online review site, along with the item itself. Although many alignment studies allow for "open" judgments of items

to any standard, we opted to restrict the review to only the standard of interest for each item (essentially requiring a [dis]confirmatory judgment). A restricted alignment review like ours, with raters evaluating the alignment between an item and a pre-specified standard, could result in a more liberal estimation of the overall alignment of an item bank. For example, if three raters independently evaluated an item and determined that, among all possible standards, it aligned best with one standard, then the resulting evidence of its alignment to that standard would perhaps be greater than in a restricted review. Furthermore, such restricted reviews are also somewhat subject to confirmation bias, in that study participants are asked to confirm the researcher's a priori theory. At the same time, restricted reviews provide less of an opportunity for an item to be rated as aligned, because there is only one opportunity for alignment (i.e., one standard).

The primary purpose of our study was to investigate the degree to which content in the items matched the content in the standards they were intended to measure (Bhola, Impara, & Buckendahl, 2003). In the context of large assessment item banks (such as the context here), misalignment is of central concern, while misclassification is of somewhat less concern. Misaligned items would be removed from consideration or revised to better align prior to being included in operational tests. In other words, we used the alignment study as part of our screening criteria to select operational items for inclusion on assessment forms.

Prior to the alignment review, all items were divided into five *review sets* within each grade (15 review sets total), with each set containing 90 items. Items selected for review were randomly assigned to one of the five sets within each grade so that each set contained approximately the same number of items written to each CCSS. Each rater then independently judged the alignment of three sets based on a linking design that allowed for the calibration of all items and raters on a common scale. Three raters reviewed each set. The rater sampling design, linking raters across sets, is displayed in Table 1.

All items were disseminated for review via a secure web-based system called the Distributed Item Review (DIR; Behavioral Research and Teaching, 2013), as displayed in Figure 1. Broadly, the DIR is an online tool for presenting test items and test forms to reviewers in a user-friendly manner so they can be evaluated for bias and sensitivity, as well as alignment to standards. Test items are uploaded into the DIR and associated with researcher-defined study questions and standards for presentation to reviewers. Data were securely stored in the DIR and then exported by researchers as text files in preparation for statistical analysis.

Rating Scale Development and Operationalization

Items were rated on a 4-point ordinal alignment scale, as follows: 0 = no alignment, 1 = vague alignment, 2 = somewhat aligned, and 3 = directly aligned. Prior to participating in the study, all raters were trained to use this scale in the alignment process through an online webinar that included examples and nonexamples of items corresponding to each of these values. To aid the critical thinking required, raters were informed that prior to assigning their rating, they should first consider the scale as (more broadly) dichotomous, in which a rating of 2 or 3 meant an item was *aligned* to the paired standard, while a rating of 0 or 1 meant *not aligned*. Then, upon making their broad alignment determination, they were asked to further refine their judgment using the 4-point scale. These ordinal rating data were used to model the latent alignment of each item, as well as calculate (and correct for) rater severity.

Rating scales often (intentionally) differ in scale modifiers based on study context (see Cox, 1980; Komorita, 1963; Matell & Jacoby, 1972). As such, we designed rater trainings to mirror alignment activities and help raters become familiar with the specific modifiers used in our 4-point scale. We also viewed rater training as an essential component of the alignment process to help ensure full understanding of the rating scale modifiers as well as the overall purpose of the study, similar to previous alignment research (Bhola et al., 2003; Herman, Webb, & Zuniga, 2007; Webb, Herman, & Webb, 2007). Finally, the scale was designed to allow for a nuanced view of alignment, without creating so fine a gradation that ratings lost meaning. For example, most middle school math CCSS have multiple parts, so it is reasonable to expect that items may match *all*, *some*, *little*, or *none* of the content required by the standard. In practical terms, items must also eventually be determined to either have, or not have, sufficient alignment with standards to be included in operational assessments—a determination based largely on the approach and acceptance criteria one adopts (Rothman et al., 2002; Webb, 2007). In our process, the 4-point scale forced each rater to make a determination as to the alignment of each item while also providing an opportunity to indicate the degree of alignment.

Analysis

The MFRM, developed by Linacre (1989), takes the general form

$$\ln\left(\frac{P_{nikj}}{P_{ni(k-1)j}}\right) = B_n - D_i - F_k - C_j, \quad (1)$$

where, P_{nikj} represents the probability that person n is rated into category k on item i by rater j. The B_n term represents the estimated location on the latent trait for person n, while D_i represents the difficulty of item i, F_k represents the Rasch-Andrich threshold for the $k-1$ item category, and C_j represents the severity of rater j.

The general MFRM model presented in (1) estimates person, item, item threshold, and rater parameters. Our analysis included item, rater, and rater threshold parameters. The shift in parameterization necessitated redefining the subscripts so that the n and i terms indexed items and raters respectively, rather than persons and items. In other words, the rows of the data matrix (subscript n) represented items, while the columns (subscript i) represented raters. Because one fewer parameter was estimated, the model was equivalent in estimation to Andrich's rating scale model (1978), defined as

$$\ln\left(\frac{P_{nik}}{P_{ni(k-1)}}\right) = B_n - D_i - F_k, \quad (2)$$

where the B_n term represents the latent alignment of item n, and rater effects are partitioned into a rater-specific severity (D_i) and a category-specific step parameter (F_k). The step parameter represents the point at which the $k-1$ and k categories are equally likely (Linacre, 2001). The F_k term, therefore, represents the point at which raters are equally likely to rate an item into either of two adjacent

Table 1. Rater Sampling Plan

Grade 6					Grade 7					Grade 8				
Set 1	Set 2	Set 3	Set 4	Set 5	Set 1	Set 2	Set 3	Set 4	Set 5	Set 1	Set 2	Set 3	Set 4	Set 5
a	a	–	–	–	–	–	–	–	–	–	–	–	–	a
–	b	b	b	–	–	–	–	–	–	–	–	–	–	–
–	–	–	c	c	c	–	–	–	–	–	–	-	–	–
–	–	–	–	–	d	d	d	–	–	–	–	-	–	–
–	–	–	–	–	–	–	e	e	e	–	–	–	–	–
–	–	–	–	–	–	–	–	–	f	f	f	–	–	–
–	–	–	–	–	–	–	–	–	–	–	g	g	g	–
h	–	–	–	–	–	–	–	–	–	–	–	–	h	h
i	i	i	–	–	–	–	–	–	–	–	–	–	–	–
–	–	j	j	j	–	–	–	–	–	–	–	–	–	–
–	–	–	–	k	k	k	–	–	–	–	–	–	–	–
–	–	–	–	–	–	l	l	l	–	–	–	–	–	–
–	–	–	–	–	–	–	–	m	m	m	–	–	–	–
–	–	–	–	–	–	–	–	–	–	n	n	n	–	–
–	–	–	–	–	–	–	–	–	–	–	–	o	o	o

Note. Each letter, a – o, represents an individual rater. Each rater is further represented by one and only one row. The table displays the overlap of raters across items so that each item was rated by at least three raters, while the raters themselves linked across item sets, allowing for the calibration of raters and items on a common scale. Each set contained 90 items.

categories. Equation (2) essentially redistributes the parameters from Equation (1) so that the "respondents" are the items themselves, while the "items" are the raters judging the alignment of each item.

Andrich's rating scale model was selected primarily for the sake of parsimony. Like all Rasch models, the rating scale model here assumes that items (in these case raters) have equal discriminations. This implies that raters are assumed to be equally capable of distinguishing between items that are and are not aligned. The model further assumes common threshold values across raters, implying that the level of alignment at which raters would move an item across categories (e.g., from a 1, *vague alignment*, to 2, *somewhat aligned*) is equivalent. With only three raters per item, it is unlikely that separate thresholds would be accurately estimated. We therefore chose the simplest model to adequately represent our observed data.

Equation 2 reflects the ratings of all items on the 4-point ordinal scale being placed on the logit scale, which is useful when evaluating the *degree* of alignment, as the logit transformation more closely represents a linear equal interval scale, provided the data fit the model (Salzberger, 2010). Degree of alignment would inform test construction, but we also needed to make decisions on which items to include on the test, which was our broader intention. The logit values can easily be transformed back to the original 4-point ordinal rating scale, which can facilitate interpretations when predetermined categorical cut points are used to form global judgments about the alignment of items to standards as a basis for inclusion in operational assessments. After the analysis, items were evaluated on the original 4-point scale with the mean ratings adjusted for rater severity. Items were then tabulated into *aligned* and *not aligned* categories using a cut-off value of 2.0 to determine which items were appropriate for operational use (provided sufficient technical characteristics) and which items should be discarded or revised. For our study, only the point estimate was used for classifying items into *aligned/not-aligned* categories. However, as discussed within the results section, the standard error of each alignment estimate was also calculated, and could be used to inform the more global item alignment decisions (e.g., item inclusion during test construction). For example, items could be deemed aligned only if the upper bound of a 95% confidence interval for the alignment estimate included the top category ("direct alignment" in our study), and the lower bound of the confidence interval *did not* include the threshold judges used to form their global judgments *vague alignment* in our study). Such a method would result in a more conservative estimate of the total number of items rated as *aligned*. It is also worth mentioning, however, that the size of the standard errors depends primarily upon the number of raters reviewing the item. Because only three raters judged the alignment of each item in our study, we anticipated the standard errors to be quite large, and this information was not weighted heavily. However, in future applications, more raters could be included and the standard errors could prove helpful in determining which items are aligned, versus those needing to be revised or discarded.

Model fit statistics. Following parameter estimation, fit statistics were calculated for both items and raters. Rasch fit statistics provide an indication of how the data fit model expectations by evaluating the residuals between model expected and observed values (Linacre, 2002; Wu & Adams, 2013). We primarily evaluated items and raters based on the unstandardized mean square outfit, but also present data on the standardized outfit, as well as standardized and unstandardized versions of the mean square infit. The infit statistic weights the residual information based on the information function so values in the middle of the latent trait continuum contribute more than those on the tails (Smith, Schumacker, & Busch, 1995). The infit statistic was developed because mean square outfit tends to be sensitive to unexpected responses at the extremes of the latent trait. For example, if a severe rater judged a "difficult to endorse" item as *aligned*, the unexpected observation would quite dramatically change the outfit value for the item (Smith et al., 1995). In their unstandardized form, both statistics center on 1.0. Evaluation of adequate fit depends on the specific study context, but is primarily driven by the sample size (Wu & Adams, 2013). In our study, each rater judged the alignment of 270 items, and we would therefore expect approximately 95% of outfit values

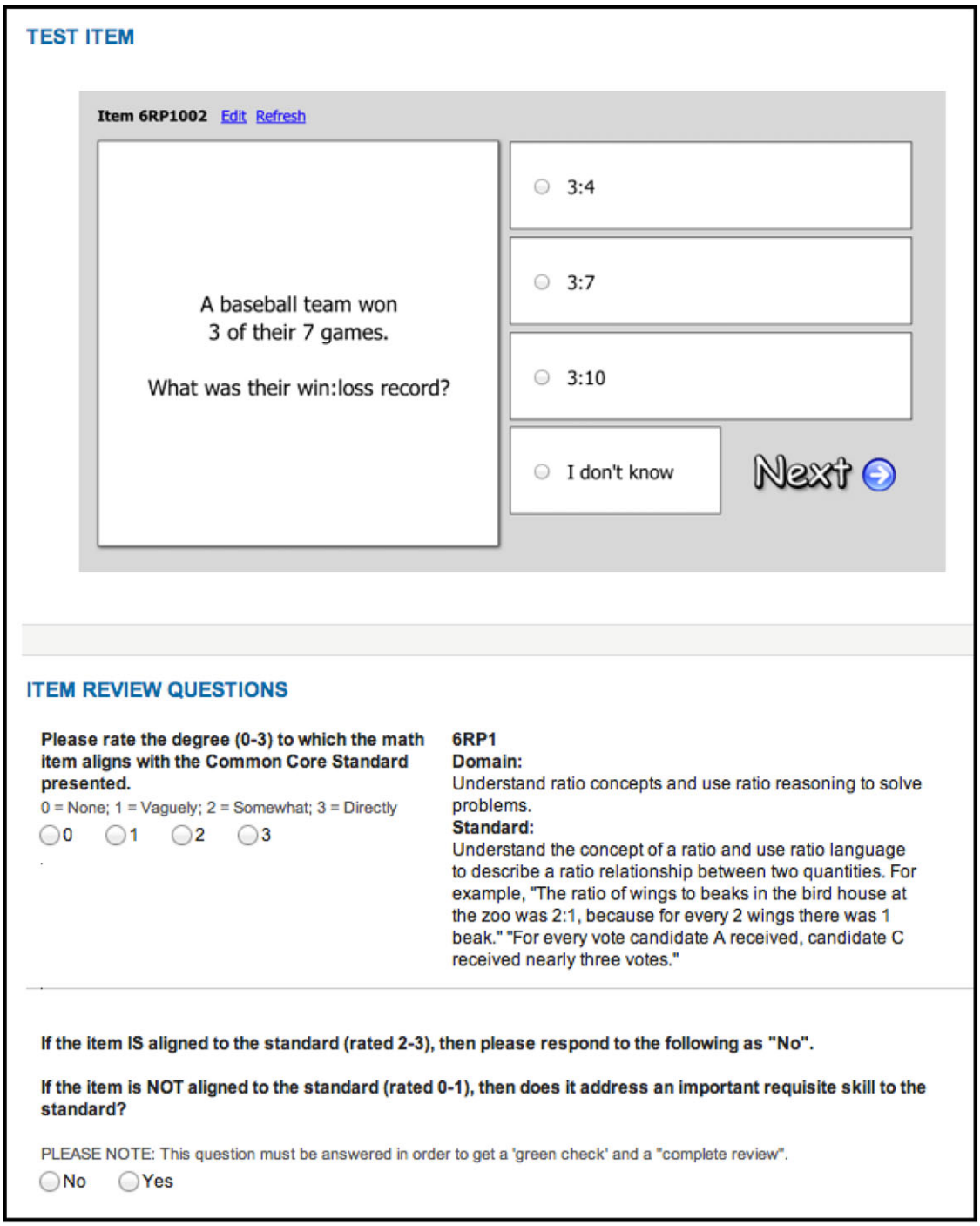

FIGURE 1. Online web-based tool used for gauging item alignment. The Distributed Item Review (DIR) is a web-based secured system for presenting test items and test forms to experts across a broad geographic region for reviews of bias, sensitivity, and alignment to standards. In this study, the DIR was used as the online framework for gauging item alignment to the CCSS.

to lie between .83 and 1.17 (see Wu & Adams, 2013). Each item was rated by only 3 raters, so we would expect item outfit values to be much more broad. We set 1.5 as our cutoff criteria for items warranting further evaluation.

The fit statistics for *raters* provide an indication of intrarater consistency and the degree to which the rater is discriminating between aligned and not aligned items. Values below 1.0 indicate less variability than expected, and a steeper rater characteristic curve (more discriminating) than expected by the model, while values above 1.0 indicate the reverse (Wu & Adams, 2013). The fit statistics for *items* can provide an indication of interrater agreement. Items that underfit the model expectations (values > 1.0) indicate an unexpected rating, conditional on the rater severity and other ratings of the item. Items with large disagreements, therefore, will generally underfit the model expectations, because the statistic is conditional on the other ratings of the item. Overfit items have less variability than is expected by the model, and often indicate high agreement among raters (i.e., all raters provide the same rating). Generally, items and raters

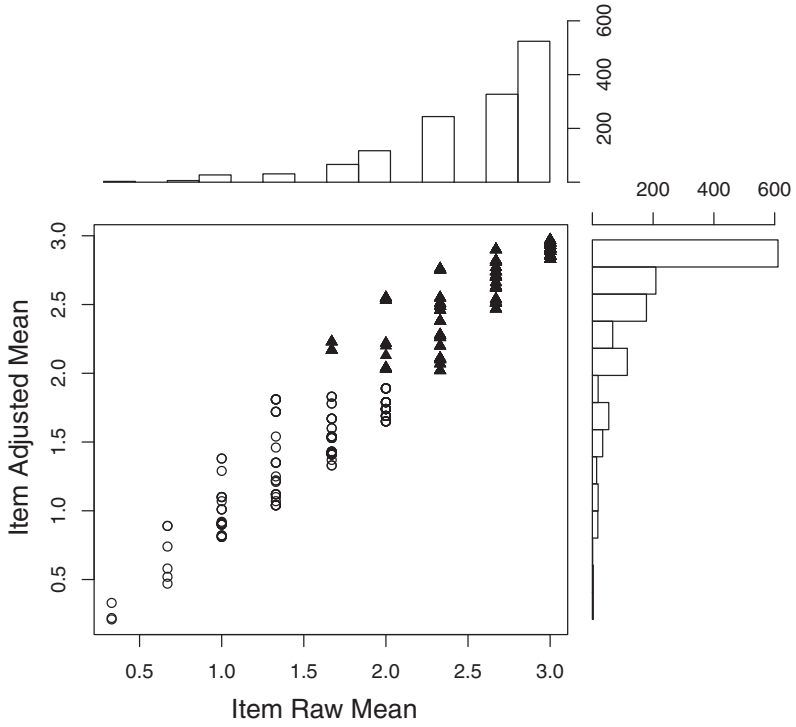

FIGURE 2. Scatterplot and distributions of Raw and rater adjusted item alignment means. Plot displays the relation between the raw and rater adjusted means for items. The item-adjusted mean corresponds to the mean item rating, conditional upon the specific raters who judged the items. Items deemed aligned (adjusted mean ≥ 2.0) are displayed with black triangles, while those deemed not aligned (adjusted mean < 2.0) are displayed with open circles. Note that for many raw mean categories there are items represented as both aligned and not aligned based on the rater adjusted mean scale. Additionally, the majority of items fell around the top end of the scale and multiple items may be represented by a single (overlapping) triangle as shown by the marginal distributions. Axes in the upper right-hand portion of the plot correspond to the frequency of items.

underfitting the model expectations are of greater concern than those that overfit the model, because they include unexpected observations and more "noise" in the data. Items and raters exhibiting extreme underfit (see Wu & Adams, 2013) may be candidates for removal. It is also worth noting that complete agreement for an item at the top or bottom of the scale for an item (i.e., all raters determine the item meets the criteria for a 0 or 3) results in fit statistics not being calculated, and "perfect" fit reported. The analysis was conducted with the FACETS software, version 3.70 (Linacre, 2012).

Results

Figure 2 displays the relation between mean item estimates on the raw and adjusted raw scales. The adjusted mean represents the nonlinear transformation from the logit scale back to the raw scale, controlling for rater severity. The distributions of items are plotted for each scale on the margins. Note that the distributions were negatively skewed, resulting in many overlapping data points on the upper end of the scale. Items rated as *aligned* (adjusted mean ≥ 2.0) are plotted with dark triangles while those rated as *not aligned* (adjusted mean < 2.0) are plotted with open circles. While the relation between the scales is quite strong ($r = .93$), there are meaningful differences from a holistic perspective. For example, roughly half the items with a raw mean of 2.0 (55%) had an adjusted mean of 2.0 or greater, after controlling for rater severity. Similarly, roughly a third of all items with a raw mean of 1.67 (32%) had an adjusted mean of 2.0 or greater, after controlling for rater severity. In other words, while the overall relation between the scales was high, the inference of alignment changes on an item-by-item basis when controlling for rater severity.

The relation between rater raw and adjusted means is displayed in Figure 3, with frequency distributions again plotted on the margins. It is worth noting that, just as item estimates are conditional on raters, rater estimates are conditional on items. The correlation between raw and adjusted means was slightly lower for raters, ($r = .81$), which is perhaps not surprising given that the items were quite diverse and not assumed to be homogenous in their alignment across item sets. The latent trait method therefore provides, in some cases, a quite different perspective on the severity of the rater (e.g., Rater 7 had an adjusted mean nearly a half point higher than the raw mean, as highlighted in Figure 3).

Item Standard Alignment

Results for 15 randomly selected items from the 1,345 items reviewed in this study are presented in Table 2. The table is intended to provide an example of the type of information obtained for the sampled item pool. For each item, observed and adjusted means are reported. The adjusted mean was used in all cases to determine whether an item did or did not align with the corresponding standard, with values below 2.0 deemed *not aligned* and therefore not appropriate for inclusion in operational test forms. A 95% confidence interval is presented along with each adjusted mean to provide an indication of the precision of the estimate. Note that because the values represent nonlinear transformations from the logit

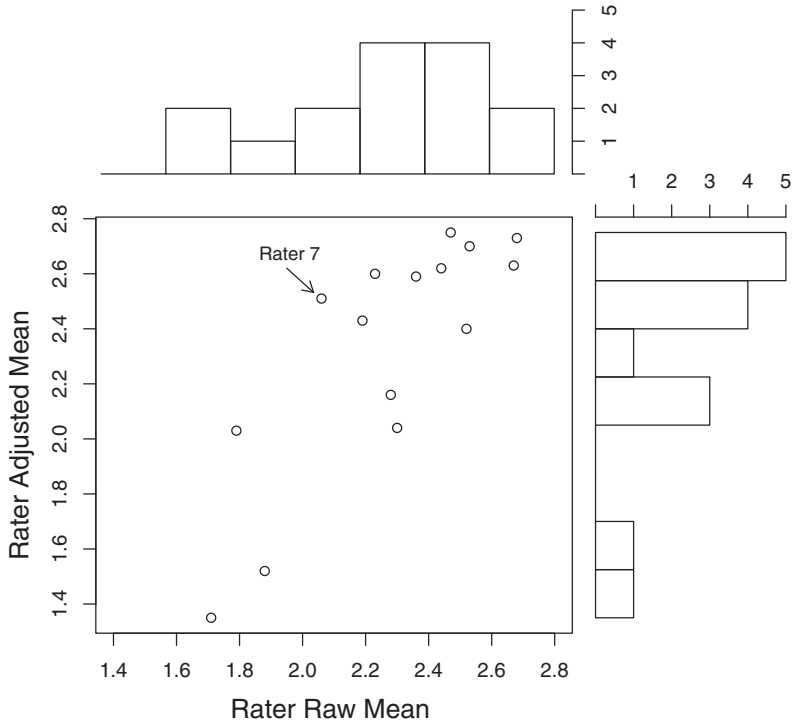

FIGURE 3. Scatterplot and distributions of raw and item adjusted rater means. Plot displays the relation between the raw and adjusted means for raters, as well as the distribution of raters on the raw and adjusted mean scales. The rater-adjusted mean corresponds to the raters' mean item rating, conditional upon the specific items the rater judged. Axes in the upper right-hand portion of the plot correspond to the frequency of raters. Rater 7 is highlighted to note the large potential differences in observed rater severity when using the raw or adjusted scales.

Table 2. Random Sample of 15 Items: Item-Standard Alignment Results

		Raw	Adj	95% CI		Logit	Fit Statistics			
Item	Grade	*M*	*M*	Lower	Upper	(*SE*)	Infit	Infit *Z*	Outfit	Outfit *Z*
1	6	2.00	1.69	.92	2.52	.50 (.75)	1.57	.98	1.65	1.07
2	6	2.00	1.74	.95	2.55	.57 (.75)	.60	−.51	.62	−.48
3	7	2.67	2.77	1.83	2.96	2.77 (1.05)	.57	−.15	.52	−.20
4	8	2.67	2.81	1.90	3.00	2.99 (1.10)	.17	−1.00	.16	−.54
5	6	2.67	2.71	1.66	2.96	2.52 (1.06)	.95	.29	.96	.34
6	8	1.67	1.83	.99	2.62	.71 (.77)	.03	−2.69	.03	−2.61
7	8	2.00	2.03	1.05	2.76	1.04 (.87)	.42	−.62	.43	−.60
8	8	3.00	2.97	2.18	3.00	4.92 (1.85)	1.00	.00	1.00	.00
9	8	2.33	2.55	1.57	2.91	2.04 (.89)	.63	−.25	.55	−.29
10	8	3.00	2.95	1.84	3.00	4.42 (1.88)	1.00	.00	1.00	.00
11	8	3.00	2.89	1.34	3.00	3.55 (1.86)	1.00	.00	1.00	.00
12	8	2.67	2.90	2.30	3.00	3.63 (1.08)	.73	.04	.55	.18
13	6	3.00	2.90	1.40	3.00	3.64 (1.85)	1.00	.00	1.00	.00
14	7	2.33	2.26	1.28	2.82	1.42 (.83)	1.00	.25	.91	.13
15	6	1.67	1.41	.73	2.29	.02 (.75)	.81	−.06	.80	−.08

Note. Raw *M* = unadjusted, raw mean of ratings across raters; Adj *M* = adjusted mean, transformed from the logit scale back to the raw scale, after controlling for rater severity; Logit = model scaled alignment rating, with higher values indicating higher alignment.

scale, the interval around the estimate is not always equal. Because only three raters judged the alignment of each item, the error bands were quite large. In fact, all but two of the items presented (Item 8 and Item 12) contained the decision cut-point within the interval. Across all 1,345 items included in this study, only 145 (10.8%) were estimated with a 95% confidence interval not crossing 2.0, with 34 (2.5%) of these being below 2.0 and 114 (8.5%) being above 2.0. Including more raters could decrease the width of the error band, leading to more items being estimated as statistically different from the decision point ($p < .05$). Following the confidence interval is the alignment estimate reported on the logit scale, along with its standard error. Higher logit values indicate more highly aligned items. Items with more extreme ratings generally had higher standard errors.

When examining the ratings of items, the raw mean is often different from the adjusted mean. For example, items 1, 2, and 7 had a raw mean of 2.0, indicating alignment for all items. When accounting for the raters who judged the items, however, the inference changes. Both items 1 and 2 had an adjusted mean below 2.0 (see Table 2), while item 7 had an adjusted mean just above 2.0. When accounting for the

Table 3. Summary of Rater Effects

Rater	Count	Raw *M*	Adj *M*	95% CI		Severity (*SE*)	Fit Statistics			
				Lower	Upper		Infit	Infit *Z*	Outfit	Outfit *Z*
1	268	2.52	2.40	2.26	2.53	−.01 (.14)	1.25	1.83	.97	−.15
2	270	2.19	2.43	2.31	2.53	−.07 (.12)	.96	−.34	1.04	.34
3	269	1.71	1.35	1.25	1.45	1.78 (.09)	.75	−3.20	.76	−2.98
4	268	2.30	2.04	1.91	2.17	.64 (.11)	.91	−.88	.87	−1.21
5	270	2.36	2.59	2.48	2.67	−.44 (.13)	1.20	1.57	1.07	.56
6	269	2.23	2.60	2.51	2.68	−.49 (.12)	1.28	2.38	1.16	1.31
7	269	2.06	2.51	2.40	2.59	−.23 (.11)	1.19	1.84	1.16	1.48
8	270	1.79	2.03	1.89	2.17	.66 (.12)	.99	−.02	.96	−.35
9	269	2.44	2.62	2.53	2.70	−.54 (.13)	.98	−.16	.92	−.48
10	269	2.53	2.70	2.61	2.76	−.80 (.15)	.78	−1.71	.92	−.41
11	269	2.68	2.73	2.65	2.77	−.92 (.15)	.87	−.84	.99	.04
12	269	1.88	1.52	1.40	1.64	1.49 (.10)	.76	−2.97	.80	−2.36
13	269	2.47	2.75	2.68	2.78	−1.00 (.14)	1.34	2.43	1.12	.82
14	268	2.67	2.63	2.51	2.71	−.55 (.15)	1.22	1.49	.89	−.55
15	269	2.28	2.16	2.02	2.29	.45 (.12)	.97	−.21	.91	−.79

Note. Count = number of items rated. The severity statistic is reported on logit scale, with higher values indicating a more severe rater.

severity of the raters, our inferences of the alignment of items 1 and 2 are reversed (from *aligned* to *not aligned*), while item 7 maintains its aligned rating. These trends are also apparent in Figure 2, as only about half of the items (55%) with a raw average of 2.0 were determined sufficiently aligned for consideration in operational test forms (dark triangles), after controlling for rater severity (i.e., using the adjusted mean scale). A few items also had a raw average below 2.0, but after controlling for rater severity, were judged adequately aligned (open circles).

After the display of the standard error, a series of item-level residual fit statistics are reported. Items that underfit the model expectations are generally more worrisome than overfit items. In the random sample of 15 items, only one (item 1) underfit the model expectations with a mean square outfit of 1.65, and a weighted mean square infit of 1.57. These values hint at a possible disagreement among raters because at least one of the ratings was unexpected, based on the other item ratings and the severity of the rater providing the unexpected rating. Using both the standard error and the fit statistics, we can begin to formulate degrees of confidence in the observed alignment rating. For items that do not fit the model well and have large standard errors, our confidence in the alignment estimate would be lower, relative to items that fit the model well and have small standard errors.

Overall, the latent trait method provided a slightly more conservative representation of item alignment. Approximately 87.7% of all items had an adjusted mean of 2.0 or more, while 90.1% had a raw mean of 2.0 or more. Approximately 5.5% of all items switched alignment categories, with approximately 3.9% shifting from *aligned* to *not aligned*, and the remaining 1.6% shifting from *not aligned* to *aligned*. It is important to recognize that when applying latent trait methods, nearly all items will shift in their alignment rating, conditional on the raters judging the item. However, only those with a raw rating on the boundary of decision points will potentially shift categories.

Approximately 9% of items with an adjusted mean above 2.0 had a mean square outfit over 1.5, indicating possible disagreement among raters. These items would likely need to be re-evaluated by another set of raters or removed from the final set of items evaluated for inclusion in operational test forms. Because only three raters judged the alignment of each item in our study, the standard errors were quite large. Only 7% of items rated as aligned did not include 2.0 within the 95% confidence interval around the adjusted mean. These results suggest additional raters should likely have been included, as the precision of estimates was low.

Rater Effects

A summary of rater effects for all 15 raters included in the study is displayed in Table 3.

As with the presentation of items, raw and adjusted means are presented, along with a 95% confidence interval around the adjusted mean. Rater adjusted means are estimated conditional on the set of items the rater judged, and represent a transformation from the logit scale (termed *severity*) back to the original 4-point rating scale. Overall, raters differed substantially in their average ratings. An examination of the adjusted mean column indicates that on average the most severe rater (Rater 3) judged items nearly one and a half categories below the most lenient rater (Rater 11), after controlling for the items raters judged. As shown in Figure 3, the ordering of rater severities differed based on whether or not the item pool was controlled for in the estimation (i.e., raw versus adjusted means).

The fit statistics in Table 3 also are important because they provide an indication of the internal consistency of raters, or intrarater reliability. That is, given the severity of the rater, as determined by his or her rating of all other items, and the endorsability of the item (logit estimate from Table 2), did the rater consistently rate items as would be expected? We found that all raters in our study fit the model expectations well, with mean square outfit statistics ranging from .76 to 1.16. These statistics also provide an indication of the degree to which raters discriminated between *aligned* and *not aligned* items, with lower values indicating higher discrimination. Note that the standard errors for raters are much smaller than for items, given that approximately 270 items were used in the calibration of each rater severity, while only three raters were used in the calibration of each item alignment.

Discussion

The primary purpose of this article was to propose the use of latent trait methods through online designs for evaluating

item alignment to standards, allowing the analyst to correct for rater severity and document intrarater reliability. The methodology was illustrated with a study exploring the alignment of formative middle school mathematics test items with the CCSS. The online methodology we described has significant practical advantages over in-person reviews, including: (1) an overall reduction in cost, (2) increased time flexibility, (3) ease of recruitment of participants across a broad geographical range, and (4) potential to review large numbers of items efficiently.

The results obtained from this study also highlight that important differences in item alignment can be observed when statistically controlling for rater effects. While it is unknown what the alignment ratings of specific items would have been under panel/discussion-based designs, we observed in some cases quite meaningful differences between the raw and adjusted means—in some cases moving the rating across our a priori alignment threshold. Had group discussion of ratings occurred, other item ratings likely would have changed as well, as raters worked to form common understandings of standards and item features prior to finalizing alignment ratings. In our study, we opted for independent reviews to guard against potential adverse effects of group dynamic, such as conformity and persuasion (Asch, 1956; Bond, 2005; Brown, 1988; Kocher & Sutter, 2005), while recognizing that some potential benefits related to expert discussion were also lost. We fit a statistical model that allowed us to (1) control for rater severity, (2) document intrarater consistency, and (3) evaluate an indicator of interrater agreement on an item-by-item basis. Whether group dynamics from panel methods generally lead to more or less accurate alignment ratings, as compared to independent ratings, remains an open question. Our proposed methodology represents an alternative to in-person panel designs, but the relative validity of each approach requires further research, including with respect to the use of latent trait methods.

Our research on alignment was conducted in a *prospective*, rather than retrospective, manner. That is, the alignment study was conducted prior to the construction of operational test forms, and the alignment results served as an initial "filter" for item inclusion. Given this focus, less attention was paid to test-level alignment, because the test forms could subsequently be constructed to meet our specifications for alignment based on the item-level information. Many methods for determining alignment focus primarily on the test level. For example, Webb (1999) discusses four summary statistics (categorical concurrence, depth-of-knowledge consistency, range-of-knowledge correspondence, and balance of representation) that provide an indication of overall test alignment. In our application, content matching between items and standards was the primary goal and information on other indicators of alignment were not collected.

Online alignment studies with latent trait methods should readily extend to any design in which raters are providing summary judgments about items on an ordinal scale. If, as with the Webb (1999) methodology, the match between the depth of knowledge required by a standard and item is of interest, and these matches were rated on an ordinal scale, then the latent trait approach should generalize. In this case, the depth of knowledge ratings would be conditioned on the rater providing the rating. If these data were collected prospectively, and information was obtained on each of Webb's alignment indicators, then test forms could be constructed such that they met each of Webb's criteria. Alternatively, if these data were collected retrospectively, test-level indicators of alignment could be calculated with the adjusted mean ratings for each criterion (conditional upon the raters providing the ratings). Categorical concurrence is perhaps the one exception where latent trait methods may not readily generalize, given that it is not contingent on raters and not judged on an ordinal scale.

Study Limitations

The small number of raters included in our study was perhaps the primary limitation of the design application. Including only three raters is essentially equivalent to estimating a students' latent ability of a theoretical construct by administering a test with only three polytomous items. Unsurprisingly, the resulting standard errors were large, with most items being within two standard errors of our cutoff value for forming global item judgments (aligned/not aligned). Although alignment of items is nearly certain to be an easier construct to measure than the latent ability of a student, our results suggest that only including three raters is still insufficient. Within the healthcare realm, research by Mallinson, Stelmack, and Velozo (2004) indicates that as few as seven items may be sufficient to obtain reliable estimates with Rasch modeling; it follows that online designs should include approximately seven raters per item, with more raters likely leading to more reliable estimates. These recommendations also align well with Webb (2007), who suggests 5–7 raters within panel settings. Part of the benefit of online designs is that it is relatively easy to include additional raters, as the cost is minimal when compared to panel designs. It is also worth noting that the extent to which one rater may influence the overall average decreases as the number of raters increases. The importance of adjusting for rater severity may therefore decrease as the number of raters increases.

On a related note, our study relied on a concurrent equating design to place all items on a common scale. Because of the low number of raters overall, the degree of overlap between raters was also small (see Table 1). The equating design therefore relied heavily on the parameter invariance assumption of Rasch modeling. Including more raters could, again, mitigate this concern as more reviewers could potentially overlap between each review set. The review sets could also be divided into smaller "chunks," which may facilitate more rater overlap between sets and less reliance on the parameter invariance assumption.

Finally, it is possible that in some cases the "extreme" rating of a single rater is the most accurate representation of the alignment of a specific item. Panel methods allow for the possibility of the group moving toward a single rater's extreme rating, while our method did not. However, in practice, it is generally more likely that outliers will conform to the group (Asch, 1956; Bond, 2005), which may be the case with tasks involving alignment. Extreme ratings are also quite possible within latent trait analyses (e.g., with severe raters each judging an item as aligned), but a single rater cannot, by design, dramatically shift the group rating.

Future Research

The reliability of measurement depends on multiple factors, but determining the number of raters sufficient for alignment

studies is a critical area of future research for online alignment designs using latent trait methods. Adding additional well-trained raters should always reduce standard errors and increase reliability, but it is likely that there is a point of diminishing returns. Where this threshold lies remains both an empirical and practical question, and may depend on the resources available for the study, the nature of the assessment being investigated (high or low stakes), and the complexity of the design (Bhola et al., 2003).

We applied Andrich's (1978) rating scale model, which assumes common, fixed threshold values across raters. This model was selected primarily for the sake of model parsimony, given the limitations of our data listed above. However, future research with more raters and more overlap among raters should explore alternative scaling methods. Specifically, Masters's partial credit model (1982), which would allow threshold values to vary across raters, as well as more general models that treat raters as a random facet varying by item (Muckle & Karabatsos, 2009), are worth further study. For example, a rater with content expertise in statistics may gauge the alignment of *statistics and probability* items more stringently than other domains. Rater severities may also vary by facets not included in our analysis, but that could be included within the general MFRM framework. Future research should investigate alternative equating designs to help determine how items can most efficiently and effectively be placed on a common scale.

Our methodology included restricted reviews, where items were paired with standards prior to dissemination. However, open reviews, in which raters first determine the standard to which an item aligns and then judge the degree of alignment, have also been conducted in previous research (e.g., Webb, 1999). Both the online and latent trait methods should readily extend to open reviews, but subtle complications may arise. For instance, standards may need to be modeled as a facet in the measurement process, as specific standards might be more or less difficult to have items align with than others, given their complexity. Future research should explore the application of latent trait methods within a variety of alignment study designs.

Although little consensus currently exists on the most appropriate mode of arriving at an alignment rating, many of the practical problems of previous alignment research could be addressed through the application of latent trait methods with online reviews—rater severity could be documented and controlled, reviewers could participate from diverse geographical areas, and a large pool of items could be efficiently reviewed. Yet, the method proposed here is not without its challenges and limitations, and future research should continue to explore the viability of different approaches of arriving at alignment ratings.

Acknowledgments

Funds for the data set used to generate this report came from a federal grant awarded to the UO from the Institute of Education Sciences, U.S. Department of Education: Developing Middle School Mathematics Progress Monitoring Measures (R324A100026 funded from June 2010 – June 2014).

References

Andrich, D. A. (1978). A rating formulation for ordered response categories. *Psychometrika*, *43*, 561–573. doi:10.1007/BF02293814.

Asch, S. E. (1956). Studies of independence and conformity: A minority of one against a unanimous majority. *Psychological Monographs: General and Applied*, *70*(9), 1–70. doi:10.1037/h0093718

Behavioral Research and Teaching. (2013). *Distributed Item Review*. Retrieved March 27, 2013, from http://www.brtitemreview.com.

Bhola, D. S., Impara, J. C., & Buckendahl, C. W. (2003). Aligning tests with states' content standards: Methods and issues.*Educational Measurement: Issues and Practice*, *22*(3), 21–29.

Bond, R. (2005). Group size and conformity. *Group Processes Intergroup Relations*, *8*, 331–354. doi:10.1177/1368430205056464

Brown, R. (1988). *Group processes: Dynamics within and between groups*. Cambridge, MA: Basil Blackwell.

Case, B. J., Jorgensen, M. A., & Zucker, S. (2004). *Alignment in educational assessment*. Pearson. Available at http://images.pearsonassessments.com/images/tmrs/tmrs_rg/AlignEdAss.pdf?WT.mc_id=TMRS_Alignment_in_Educational_Assessment

Council of Chief State School Officers (CCSSO). (2002). *Models for alignment analysis and assistance to states*. Washington, DC: Author. Available at http://images.pearsonassessments.com/images/tmrs/tmrs_rg/AlignEdAss.pdf?WT.mc_id=TMRS_Alignment_in_Educational_Assessment

Cox, E. P. (1980). The optimal number of response alternatives for a scale: A review. *Journal of Marketing Research*, *17*, 407–442.

Engelhard, G. (1994). Examining rater errors in the assessment of written composition with a many-faceted Rasch model. *Journal of Educational Measurement*, *31*, 93–112.

Herman, J. L., Webb, N. M., & Zuniga, S. A. (2007). Measurement issues in the alignment of standards and assessments: A case study. *Applied Measurement in Education*, *20*(1), 101–126.

Kocher, M. G., & Sutter, M. (2005). The decision maker matters: Individual versus group behaviour in experimental beauty contest games. *The Economic Journal*, *115*, 200–223.

Komorita, S. S. (1963). Attitude content, intensity and the neutral point on a Likert scale. *Journal of Social Psychology*, *61*, 327–334.

La Marca, P. M. (2001). Alignment of standards and assessments as an accountability criterion. *Practical Assessment, Research and Evaluation 7*(21). Retrieved October 17, 2012, from http://pareonline.net/getvn.asp?v=7&n=21

Laporte, L., van Nimwegen, C., & Uyttendaele, A. J. (2010, October) *Do people say what they think? Social conformity behavior in varying degrees of online social presence*.Paper presented at the proceedings of the 6th Nordic Conference on Human-Computer Interaction: Extending Boundaries, Reykjavik, Iceland.

Linacre, J. M. (1989). *Many-facet Rasch measurement*. Chicago, IL: MESA Press.

Linacre, J. M. (2001). Category, step and threshold: Definitions and disordering. *Rasch Measurement Transactions*, *15*, 794.

Linacre, J. M. (2002). What do infit and outfit, mean-square and standardized mean? *Rasch Measurement Transactions*, *16*, 878.

Linacre, J. M. (2012). *Facets computer program for man-facet Rasch measurement (Version 3.70.0)*. Beaverton, OR: Winsteps.com.

Mallinson, T., Stelmack, J., & Velozo, C. (2004). A comparison of the separation ratio and coefficient alpha in the creation of minimum item sets. *Medical Care*, *42*, 117–124. doi:10.1097/01.mlr.0000103522.78233

Martone, A., & Sireci, S. G. (2009). Evaluating alignment between curriculum, assessment, and instruction. *Review of Educational Research*, *79*, 1332–1361. doi:10.3102/0034654309341375

Masters, G. N. (1982). A Rasch model for partial credit scoring. *Psychometrika*, *47*, 149–174. doi:10.1007/BF02296272

Matell, M. S., & Jacoby, J. (1972). Is there an optimal number of alternatives for Likert scale items? Effects of testing time and scale properties. *Journal of Applied Psychology*, *56*, 506–509.

Muckle, T. J., & Karabatsos, G. (2009). Hierarchical generalized linear models for the analysis of judge ratings. *Journal of Educational Measurement*, *46*, 198–219.

Myford, C. M., & Wolfe, E. W. (2000). *Strengthening the ties that bind: Improving the linking network in sparsely connected rating designs* (TOEFL Technical Report No. 15). Princeton, NJ: Educational Testing Service.

No Child Left Behind Act, Pub. L. No. 107–110 (2001).
Porter, A. C. (2002). Measuring the content of instruction: Uses in research and Practice. *Educational Researcher*, *31*, 3–14. doi:10.3102/0013189×031007003
Porter, A. C. (2004). Curriculum assessment. In J. Green, B. Camilli, & P. Elmore (Eds.), *Complementary methods for research in education*. Washington, DC: American Educational Research Association.
Rothman, R., Slattery, J. B., Vranek, J. L., & Resnick, L. B. (2002). *Benchmarking and alignment of standards and testing* (Technical Report 566). Washington, DC: Center for the Study of Evaluation.
Salzberger, T. (2010). Does the Rasch model convert an ordinal scale into an interval scale? *Rasch Measurement Transactions*, *24*, 1273–1275.
Smarter Balanced Assessment Consortium. (2012). *Computer adaptive testing*. Retrieved December 5, 2012, from http://www.smarterbalanced.org/smarter-balanced-assessments/computer-adaptive-testing/.
Smith, R. M., Schumacker, R. E., & Busch, M. J. (1995, April). *Using item mean squares to evaluate fit to the Rasch model*. Paper presented at the Annual meeting of the American Educational Research Association, San Francisco.
Webb, N. L. (1999). *Alignment of science and mathematics standards and assessments in four states* (Research Monograph No. 18). Madison, WI: University of Wisconsin-Madison, National Institute for Science Education.
Webb, N. L. (2007). Issues related to judging the alignment of curriculum standards and assessments. *Applied Measurement in Education*, *20*, 7–25. doi:10.1080/08957340709336728
Webb, N. L., Alt, M., Ely, R., & Vesperman, B. (2005). *Web Alignment Tool (WAT): Training manual (Version 1.1)*. Council of Chief State School Officers, Wisconsin Center for Education Research. Available at http://wat.wceruw.org/Training%20Manual%202.1%20Draft%20091205.doc
Webb, N. M., Herman, J. L., & Webb, N. L. (2007). Alignment of mathematics state-level standards and assessments: The role of reviewer agreement. *Educational Measurement: Issues and Practice*, *26*(2), 17–29. doi:10.1111/j.1745–3992.2007.00091.x
Wolfe, E. W. (2004). Identifying rater effects using latent trait models. *Psychology Science*, *46*, 35–51.
Wu, M., & Adams, R. J. (2013). Properties of Rasch residual fit statistics. *Journal of Applied Measurement*, *14*, 339–355.