Using a Model of Analysts' Judgments to Augment an Item Calibration Process

Educational and Psychological

Measurement

2015, Vol. 75(5) 826–849

© The Author(s) 2014

Reprints and permissions:

sagepub.com/journalsPermissions.nav

DOI: 10.1177/0013164414559775

epm.sagepub.com

**Carl Hauser¹, Yeow Meng Thum¹, Wei He¹,
and Lingling Ma^{1,2}**

Abstract

When conducting item reviews, analysts evaluate an array of statistical and graphical information to assess the fit of a field test (FT) item to an item response theory model. The process can be tedious, particularly when the number of human reviews (HR) to be completed is large. Furthermore, such a process leads to decisions that are susceptible to human errors. A key finding from behavioral decision-making research has shown that a parametric model of human decision making often outperforms the decision maker himself. We exploit this finding by seeking a model to mimic how analysts integrate FT item level statistics and graphical performance plots to predict the analyst's assignment of the item's status. The procedure suggests a set of rules that achieves a desired level of classification accuracy, separating situations in which the evidence supports firm decisions from those situations that would likely benefit from HRs. Implementation of the decision rules accounts for an estimated 65% reduction in calibrations requiring HRs.

Keywords

item calibration, item fit statistics, human decision making, logistic regression, item review

Good item parameter estimates are critical to the validity of a test based on item response theory (IRT). Approaches to assessing the quality of estimates obtained during an item calibration process have been a subject of numerous studies for well over

¹Northwest Evaluation Association, Portland, OR, USA

²USAA, San Antonio, TX, USA

Corresponding Author:

Carl Hauser, Northwest Evaluation Association, 121 NW Everett Street, Portland, OR 97209, USA

Email: carl.hauser@nwea.org

three decades (e.g., Drasgow, Levine, Tsien, Williams, & Mead, 1995; Glas & Suarez-Falcon, 2003; Hambleton & Han, 2005; Orlando & Thissen, 2000, 2003; Smith, 2000; Stone & Zhang, 2003). These investigations have provided a range of tools for ensuring defensible psychometric functioning of items and tests. Among these tools are several test-item descriptive statistics, the results of several tests of item fit in common use, as well as graphical plots of an item's observed responses or residuals from the theoretical model used to calibrate the item (e.g., Hambleton & Han, 2005; Hambleton, Swaminathan, & Rogers, 1991). Various descriptive statistics (e.g., percent correct, point-measure correlation) are employed as part of evaluating an item's calibration because each provides a slightly different aspect of the item's behavior within the sample of students that responded to it. Furthermore, item fit statistics (e.g., infit, outfit, standardized χ^2) provide useful information about how well the responses adhere to the expectations of the IRT model. Ideally such information about an item would make up the evidence base used in a human review (HR) to support decisions about the item's operational status.

While the forms of information collected about an item's calibration status are empirical, objective, and repeatable, they do not by themselves result in a decision. For example, it is not always clear when item fit is adequate in any one case. What numerical value constitutes poor fit? When some indicators suggest good quality but others suggest caution, what should drive the decision? Clearly, HRs are subjective; even though the information being considered is from objective sources, their confluence in the form of a summary of field test (FT) performance can include ambiguous signals that lack defined rules upon which to act. Recognizing this, the integration of the available item calibration information involves the analyst's judgment and will, in varying degrees, reflect the analyst's knowledge and experience. How stable should we expect an analyst's judgments to be? Would the decision made by another analyst be the same?

Best practice would seem to suggest that all FT items passing through the calibration process should undergo review by at least two experienced analysts. Judgments from several analysts would guard against an overreliance on the judgments of any one analyst. However, HRs can be expensive and inefficient, especially when large numbers of items are under consideration. They consume staff time and divert staff attentions away from other priorities. To the extent that HRs can be reduced without substantial increase in the mischaracterization of item status, the more efficient the calibration process will be.

Consider how issues may arise in the case of item calibration for a computerized adaptive test (CAT). Assume that the field-testing procedure includes embedding two to three FT items in operational tests and presenting these items to examinees as part of each normal test administration. For each FT item, the "calibration engine" monitors the number of responses and computes a calibration after a threshold number of responses has been reached. Once a calibration has been computed, a report is generated that provides fit and descriptive statistics and a plot of item responses cast against the item's theoretical response curve. These reports are presented to analysts

to examine the empirical item response curves and accompanying psychometric indicators. Based on this information, analysts make determinations about how the item should be considered going forward. As indicated above, this process can be both time-consuming and subjective. Decisions on any given item may vary among analysts. Moreover, an individual analyst may render different decisions on a given item if it is presented on different occasions.

While response plots and the accompanying psychometric indicators are intended to serve as the primary focus, little is known about how those indicators are used in shaping the analyst's decision. The absence of studies examining the accuracy and stability of human judgments about FT item performance or of the role such judgments play in the evaluation of items suggests several plausible explanations. One is that human review is considered as simply a perfunctory exercise tacked onto an item calibration and evaluation process; that is, it holds little sway in the decision to make an item operational. Another possibility is that decisions about items are based purely on statistical analyses of item performance data. Or it may be that the process of human review of FT performance data is felt to be so canonically understood that little would be gained by studying it further, or at all. Yet another possibility is that the role of inherently subjective human judgment inside an otherwise objective process, if not antithetical to the goals of the process, is at least difficult to reconcile. We suspect that there is some truth to each of these explanations and probably others.

The indifference to studying human judgments made in the context of professional practice is not unique to this area of psychometrics. Indeed, a similar aversion has been present in clinical psychology for roughly eight decades (Meehl, 1954). Since Meehl's seminal book, the value and in fact, the superiority of using actuarial or statistical models versus clinical methods to make judgments or predictions has been demonstrated convincingly (e.g., Dawes, 1979; Dawes, Faust, & Meehl, 1989; Einhorn, 1974; Goldberg, 1968). However, the recognition, acknowledgment, and acceptance of this evidence has been mixed (Dawes et al., 1989; Swets, Dawes, & Monahan, 2000). Other fields in which "clinical" judgment based on multiple sources of objective data and information play a major role, such as personnel selection, criminology (e.g., predicting parole violations, recidivism), medicine (e.g., diagnosis, radiology, oncology), and education (e.g., predicting college success), have all shown similar findings (Dawes, 1993; Dawes et al., 1989; Swets et al., 2000). The overwhelming finding that actuarial models of human decision making consistently outperform clinical decision makers holds important implications for the practice of making judgments about FT item performance. There is little reason to believe that psychometricians or other trained analysts reviewing item calibration reports to make judgments about items' operational status going forward would somehow be immune to the problems of judgment consistency and accuracy that are so pervasive among professionals in other fields over the past 50 years (Dawes et al., 1989). Dawes (1979) has contended that people, especially experts, are much better at selecting and coding information than they are at integrating it, as in judgment tasks.

We argue here that the processes of making what are essentially “clinical” judgments about FT item performance will be well served by studying, if not their “weakest” aspect, then the aspect that has been virtually unaddressed.

More specifically, we wanted to know the extent to which analysts’ judgments of FT items’ status can be predicted from the statistical and graphical information used to make those judgments. From this motivation, four more narrowly focused questions emerge. They are:

1. To what extent does a model of the analyst replicate the analyst’s actual decisions?
2. To what extent does prediction accuracy vary between item subject areas?
3. What criteria would be required to minimize serious classification errors that would offset the benefits of using the prediction model in practice?
4. What changes in efficiency would be expected by embedding a model of analysts’ judgments into an item review process in which all items considered to be successfully calibrated undergo HR?

The importance of the current study lies in its *initial* focus on this point, namely, that a model of how analysts use data and graphical displays of item performance in judging items’ subsequent status may be used to improve the quality and efficiency of the calibration process.

Our approach to understanding how analysts employ item information in deciding an item’s subsequent status can be traced to well-established findings that a prediction model, based on an expert’s behaviors, tend to outperform the expert himself (viz., Armstrong, 2001; Dawes, 1979; Dawes et al., 1989; Einhorn, 1974; Goldberg, 1968; Libby, 1976; Weiss, Shanteau, & Harries, 2006). In this approach predictive models are developed that effectively serve as substitutes for the decision maker’s actual judgments. Following this line of inquiry, a model of the expert is built in the first phase of the study from the data used by analysts to form evaluative decisions about items.

In the second phase of the study, decision meta-rules based on the predictive model are applied *in place of the expert*. These rules were “tuned” for the model’s operating characteristics. The expectation is that these rules, when integrated into the broader flow of the item calibration process, should improve the accuracy and consistency of judgments of FT item calibration status.

It should be understood from the outset that the focus of the study was on the judgments of the specific analysts who reviewed the item calibration reports. This, of course, limits generalizability of the results to the future judgments of these analysts when they judge similar (dichotomously scored, multiple choice) items. We would argue, however, that the steps undertaken in the methodology would be expected to generalize to other similar contexts and situations, that is, large-scale item calibration efforts that include human review as a key component in determining the fate of calibrated test items. Item types may differ, and the particular statistical and graphical

information provided to analysts may vary, but the analytic procedures can be readily adopted from the methods below.

Method

Calibration Records

An extant set of item calibration records that had been reviewed by two experienced analysts over a recent 14-month period was used as the primary dataset. All records were generated from responses to field test items embedded in operational computerized adaptive tests (CATs) developed to assess K–10 student achievement in mathematics, reading, language usage, and general science. Each calibration record contained a set of statistics related to the item's characteristics along with a plot of test takers' responses to the item. The plot included the theoretical ogive corresponding to the dichotomous Rasch model plus the response plot for each answer option. Analyst used the combination of the graphical plot and accompanying statistics to make a trichotomous judgment about the item: Accept (make operational), reject (remove from further consideration or make revisions before additional field-testing), or return the item to the FT status. How the analysts treated the available information was not strictly prescribed. They were asked to simply make a decision using both the visual and item fit information, guided by commonly employed thresholds for the fit statistics. Past records showed that roughly 45% of all items were returned to FT status at least one time, resulting in multiple calibration reports for a substantial proportion of items. For example, 10% of reading items were reviewed four to five times and 10% of the mathematics items were reviewed seven to eight times.

The need to validate the prediction model with a dataset composed of completely different items necessitated splitting the full dataset. Distinct *items* ($N = 8,017$) were randomly assigned to one of two groups: (a) a "training" group ($n = 5,292$; 66%) and (b) an evaluation group ($n = 2,725$; 34%). For the training group, only the item's last cycle through the calibration/HR process was used. These records were considered to represent the most complete information available for making a decision about the status of a FT item. For the evaluation group, a single instantiation of the item's presence in the calibration/HR process was randomly selected. These records were considered to represent those situations a reviewer is most likely to encounter. Table 1 provides a summary of how the reviews for the items broke out in each group in terms of the scale (subject) and HR decision.

Each calibration record included a number of variables related to the item and its performance. The presence of these variables in the calibration record made them accessible to analysts to use in their evaluations of items. The variables, identified as "base" in Table 2, were used either as potential predictors or outcomes. If necessary, base variables were transformed to make them more suitable for the analyses. The first six entries in Table 2 are fit and fit-related statistics that were attached to each calibration record. Variable numbers 7 through 12 were simple descriptive variables the analysts reported as being useful. The final three predictor variables were

Table 1. Item Reviews in Training and Evaluation Sets by Subject and Human Review Decision.

Subject	Training				Evaluation			
		Reject	Ret to FT	Accept	Total	Reject	Ret to FT	Accept
Mathematics	n	909	153	1,586	2,648	465	98	842
	Set %	34.3	5.8	59.9		33.1	7.0	59.9
Reading	n	669	33	811	1,513	375	19	395
	Set %	44.2	2.2	53.6		47.5	2.4	50.1
Lang. usage	n	240	40	320	600	104	17	144
	Set %	40.0	6.7	53.3		39.3	6.4	54.3
Science	n	216	26	289	531	124	13	129
	Set %	40.7	4.9	54.4		46.6	4.9	48.5
Total		2,034	252	3,006	5,292	1,068	147	1,510
Overall %		38.4	4.8	56.8		39.2	5.4	55.4
								2,725

Table 2. Variables Used in Logistic Regression.

No.	Base	Transform	Description
Potential predictor			
1	Infit	Infit_Rcp	Reciprocal value of the statistic commonly used to assess fit to the Rasch model based on an information-weighted chi-square. In this use, values closest to 1 represent better fit.
2	Outfit	Outfit_Rcp	Reciprocal value of the statistic commonly used to assess fit to the Rasch model based on an outlier sensitive (unweighted) conventional chi-square statistic. In this use, values closest to 1 represent better fit.
3	MSE	—	Mean-square error: the average squared difference between the percentage of correct responses expected and the percentage of correct responses observed conditioned on examinee scale score. Values closer to zero indicate better fit.
4	r_{oe}	—	Correlation of observed correct responses to Rasch model expected responses. Higher values are generally preferred, but can be deceptive since differences between observed and expected location are not accounted for.
5	Rpbim	—	Point-measure correlation: the correlation between the item response and the respondent's final RIT score. Values higher than .25 are preferred.
6	zChiSqr	—	Standardized χ^2 of observed responses to expected responses: values less than 1 are associated with better fit, but large Ns can inflate values.
7	Audio	—	Item is presented with audio (dummy coded)
8	CalQ	—	Quartile of the calibration sample that contains the item's calibration: calibrations falling in Q1 or Q4 can be a sign that the "precalibration difficulty" assigned by the content specialist was an overestimate (Q1) or an underestimate (Q4).

(continued)

Table 2. (continued)

No.	Base	Transform	Description
9	TotCnt	CntCat	Category (of total item responses): 1 (1,000–1,500), 2 (1,501–2,000), 3 (2,001–2,500) 4 (2,501–3,000) 5 (3,000+)
10	CQpos	—	Proportion of responses in the quartile bounding the calibration: indicates how far into Q1 or Q4 a calibration falls from the outside extreme of the quartile. Values close to 1 reduce concern that the calibration is not supported by an adequate range of observed data.
11	Pcorr	Pcorr_Ctr50	Proportion of correct responses (centered at .50)
12	Q1pct	—	Proportion of correct responses in quartile 1: higher values ($> .50$) can suggest poor item targeting.
13	Rpbim & Inf	Rpbim_Inf	Ratio of Point-measure correlation to the absolute value of Information-weighted fit–higher ratios suggest better fit.
14	Q1-Q4	Trend	Linear trend based on proportions of responses in quartiles: $-.15 * Q1 + -.05 * Q2 + .05 * Q3 + .15$
15	TotCnt	T_Sqrt	* Q4—positive departures from 0 indicate increasingly linear trends. Square root of total responses (truncated to 80): adopted to reduce the variance of total responses as a predictor.
Outcome			
16	HR_Accept	—	HR recoded to 1 = “Accept” or 0 = “Other (not Accept)”
17	HR_Reject	—	HR recoded to 1 = “Reject” or 0 = “Other (not Reject)”

constructed from some of the preceding variables. Dichotomous outcome variables were constructed from the status assigned by the human review process. These are included as the last two variables in Table 2.

Modeling the Analyst's Decision Behavior

For the training set data, models were built for each subject area (mathematics, reading, language usage, and science) by regressing each of the dichotomously coded outcome variables (HR_Accept and HR_Reject) onto the set of predictors using a separate logistic regression procedure with stepwise selection for each outcome. The significance criterion for a variable to enter the model was set at .50, while the significance criterion for remaining in the model was set at .10. Individual probabilities for each outcome for each item were cross-validated using the leave-one-out procedure (Shtatland, Kleinman, & Cain, 2004) as implemented in SAS. *Receiver operating characteristic* (ROC) curve analyses were performed to judge the quality of prediction for each predictor set for each outcome variable. The focus for these analyses was on identifying the probability associated with decisions of high specificity (and thus, low false positive values) to help inform the rules used to govern the classification of items using predicted probabilities. To “score” the corresponding evaluation dataset, the estimated regression coefficients $\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2, \dots, \hat{\beta}_k$, for the k predictors $x_1, x_2, \dots, x_k$ were applied to item i . For each calibration decision status (i.e., HR_Accept and HR_Reject) a logistic function was used to estimate the probability, p_i , that the decision status was present for item i . The predicted probability for either decision status may be expressed as

$$p_i = \Pr(\text{Status}_i = 1 | x_{1i}, x_{2i}, \dots, x_{ki}) = \frac{1}{1 + e^{-(\hat{\beta}_0 + \hat{\beta}_1 x_{1i} + \hat{\beta}_2 x_{2i} + \dots + \hat{\beta}_k x_{ki})}}. \quad (1)$$

Note that the set of predictors was allowed to vary by subject and HR status.

Classification Rules and Impact

Central to the analysis of the scored datasets was the treatment of the sets of probabilities coming from the two logistic regressions for each item. It was necessary to establish a set of rules to isolate probabilities from the “Accept” events (HR_Accept/HR_NOT Accept) and from the “Reject” events (HR_Reject/HR_NOT Reject) that would be used to define the predicted status for the item. The first method was based on a very simple rule for the probabilities; specifically, identifying the cell that contained the only probability of accepting (vs. not accepting), or rejecting (vs. not rejecting) the item that exceeded .5. This rule, referred to as the *base decision* method, is laid out in section A of Table 3.

A second method is being referred to as the *optimal probability* method. In this method, probabilities were isolated by using the information from the ROC curve-based classification tables (presented in the Results section, below). Specifically,

Table 3. Assigning Predicted Status Using the Rules Applied for Each Focus on Probability.

Mathematics			Reading	
Pr(HR_Accept)			Pr(HR_Accept)	
Pr(HR_Reject)	$\leq .5$	A. Base Probabilities	$\leq .5$	$\leq .5$
$\leq .5$	Return to FT	Pr(HR_Reject)	Return to FT	Return to FT
$> .5$	Reject	$\leq .5$	Reject	Reject
		$> .5$		
		B. Optimal Probabilities		
$\leq .30$	$< .55$		$< .55$	$\geq .55$
$> .30$	Return to FT		Return to FT	Accept
	Reject		Reject	Return to FT
		C. Maximum Specificity		
$.69 > P$	$P \geq .80$		$P \geq .82$	$.67 > P$
$.84 > P \geq .69$	Accept		Accept	Return to FT
$P \geq .84$	Return to FT	$.70 > P$	Return to FT	Reject w/Rev
	Return to FT	$.85 > P \geq .70$	Return to FT	Reject
		$P \geq .85$		

Youden's Index (as cited in Schisterman, Perkins, Liu, & Bondell, 2005) was adopted for this method. The sensitivity and specificity values from the analysis of a single event type (Accept or Reject) were added to determine the probability that represented optimal performance in terms of "true positives"—correct predictions of the observed event (Accept or Reject) in combination with "true negatives"—correct predictions of the nonevent (not-Accept or not-Reject). The probabilities isolated from the two event types define the cells in the mathematics and reading prediction grids shown in section B of Table 3.

Finally, the third method for isolating probabilities is being referred to as the *maximum specificity* method. This method was guided by a judgment that predictions resulting in possible serious misclassification (i.e., predicting "accept" for an item that human reviewers would "reject") are errors that outweigh the importance of "correct" classifications according to the model (Swets et al., 2000). Under this method, probabilities were also identified using the ROC curve-based classification tables (presented in Results section, below). For each event type, a primary probability level was isolated by using a high percentage of specificity (i.e., ratio of true negative to the sum of false positives and true negatives). Since high specificity is negatively related to the rate of false positives, misclassification errors would be less likely. However, as specificity percentages increase, the percentages of correct classifications decrease, potentially to the point of negating any advantages of using the procedure. To offset the negative effects of low percentage of correct classification, an *arbitrary* minimum of the percentage of correct classifications was adopted. Specifically, the primary isolated probability was the one associated with maximum specificity *and* a percentage of correct classifications that was within 5 percentage points of the maximum of percentage of correct classifications.

At the same time, the rules were developed to acknowledge the possibilities for misclassification by building in two "hedge" status categories. These categories were established between the "reject" status and the "return to FT" status as well as between the "accept" status and the "return to FT" status by arbitrarily subtracting 15 percentage points from the primary isolated probability. The primary probabilities isolated from the two event types as well as the probabilities adopted to establish the "hedge" categories define the cells in the mathematics and reading prediction grids shown in section C of Table 3.

Results

Beginning in this section, attention is predominantly limited to mathematics. Where it is deemed informative, results from reading items are also presented. It is worth noting, however, that corresponding results for reading, language usage, and science items follow very similar patterns to those observed for mathematics items. For example, descriptive statistics revealed that the training sets (one for Accept and one for Reject for each subject area) were quite comparable in values across most predictors. However, some differences were evident for predictors based on calibration

sample size, namely T_{sqrt} , CntCat , and $z\text{ChiSqr}$. For example, the mean number of items used to calibrate items in reading was about 850 larger than that used for calibration in mathematics. Distributions and inter-correlations of these predictors for all FT items are available from the authors.

Logistic Regression Analyses

Logistic regressions were carried out separately for each subject area. The first portion of this section presents summaries of the stepwise analyses in the form of maximum likelihood (ML) estimates and their standard errors (SE) for the predictors included in the mathematics analysis. The ROC curve analyses based on the final step of the regression analyses are also provided. The second portion of this section presents tables of hit rates and confusion matrices from both the model building sample, as well as from the evaluation sample.

Table 4 provides the ML estimates and their SEs for predictors included in the two training models for mathematics. Across both models, eight predictors were judged as marginal contributors under conventional significance levels (i.e., $p < .05$) and were dropped from the final models. These included the correlation of the observed response to the expected response (r_{oe}), the percent of correct responses in the first quartile ($Q1\text{pct}$), linear trend (Trend), mean square error (MSE), proportion of correct responses (Pcorr_Ctr50), category of total item responses (CntCat), square root of total item responses (T_{sqrt}), and proportion of responses in the quartile bounding the calibration (CQpos) for the Reject model only. The lower portion of Table 4 provides summaries of the ROC curve analyses for the mathematics training set. This table reveals that both models were quite accurate (minimum $\text{AUC} > .94$) at predicting a target event (e.g., the item was Accept vs. not-Accept, or the item was Reject vs. not-Reject). The reader is reminded that the area under the ROC curve (AUC) is a reflection of prediction accuracy. Typical AUC values range from .5 (flip of a coin accuracy) to 1.0 (perfect accuracy). Convention holds that $\text{AUC} > .90$ indicates high accuracy.

Estimates of each item's probabilities of Accept vs. not-Accept or of Reject vs. not-Reject were cross-validated using a leave-one-out procedure (Shatland et al., 2004) as part of the prediction model building. This procedure is tantamount to leaving out the current observation and reestimating the parameters. Thus, for each item for each model, two probabilities estimates (one from the model and one from the model estimated without the current observation) were summarized for each outcome (the event and the nonevent). The mean differences between the probabilities computed with and without the leave-one-out procedure differed only at the fifth decimal place for all outcomes except for the Accept outcome for reading items where differences appeared at the fourth decimal place.

The classification tables resulting from the ROC curve analyses for the two mathematics training models are presented together in Table 5. Classifications are based on the probabilities yielded by the cross-validation procedures. The bolded cell entries

Table 4. Maximum Likelihood Estimates and Standard Errors for the MATHEMATICS Training Models.

Parameter	Accept vs. Not Accept			Reject vs. Not Reject		
	Estimate	SE	$P > \chi^2$	Parameter	Estimate	SE
Intercept	-100.500	6.731	<.0001	Intercept	108.500	6.643
CQpos	1.381	0.721	0.055	CalQ	-0.353	0.088
CalQ	0.689	0.087	<.0001	Infit_Rcp	-126.200	8.379
Infit_Rcp	108.000	8.009	<.0001	Outfit_Rcp	7.684	2.099
Outfit_Rcp	-6.472	1.924	0.001	Rpbim	296.700	18.819
Rpbim	-274.700	18.989	<.0001	Pcorr_Ctr50	10.520	1.088
Pcorr_Ctr50	-7.266	1.371	<.0001	zCHISQR	0.381	0.038
zCHISQR	-0.330	0.039	<.0001	Rpbim_InF	-308.400	19.116
Rpbim_InF	302.000	19.804	<.0001	Audio	0.319	0.103
Audio	-0.518	0.098	<.0001			
ROC curve statistics						
Area	SE	95% Wald confidence limits	Area	SE	95% Wald confidence limits	
0.9488	0.00423	0.9405	0.9571	0.9566	0.00376	0.9492
						0.9640

Table 5. Classification Tables for the MATHEMATICS Training Models.

Prob. level	Correct		Incorrect		Percentages					Youden Index	
	Non-event		Non-event		Correct	Sensitivity	Specificity	False-POS	False-NEG		
0.30	1,530	779	283	56	Accept vs. Not Accept	87.2	96.5	73.4	15.6	6.7	0.699
0.35	1,515	821	241	71		88.2	95.5	77.3	13.7	8.0	0.728
0.40	1,496	845	217	90		88.4	94.3	79.6	12.7	9.6	0.739
0.45	1,477	868	194	109		88.6	93.1	81.7	11.6	11.2	0.748
0.50	1,456	882	180	130		88.3	91.8	83.1	11.0	12.8	0.749
0.55	1,440	903	159	146		88.5	90.8	85.0	9.9	13.9	0.758
0.60	1,411	916	146	175		87.9	89.0	86.3	9.4	16.0	0.753
0.65	1,375	940	122	211		87.4	86.7	88.5	8.1	18.3	0.752
0.70	1,337	952	110	249		86.4	84.3	89.6	7.6	20.7	0.739
0.75	1,289	964	98	297		85.1	81.3	90.8	7.1	23.6	0.721
0.80	1,231	980	82	355	83.5	77.6	92.3	6.2	26.6	0.699	
0.85	1,156	994	68	430	81.2	72.9	93.6	5.6	30.2	0.665	
0.90	1,050	1,017	45	536	78.1	66.2	95.8	4.1	34.5	0.620	
0.95	787	1,047	15	799	69.3	49.6	98.6	1.9	43.3	0.482	
					Reject vs. Not Reject						
0.30	829	1,515	224	80	88.5	91.2	87.1	21.3	5.0	0.783	
0.35	806	1,550	189	103	89.0	88.7	89.1	19.0	6.2	0.778	
0.40	787	1,579	160	122	89.4	86.6	90.8	16.9	7.2	0.774	
0.45	771	1,594	145	138	89.3	84.8	91.7	15.8	8.0	0.765	

(continued)

Table 5. (continued)

Prob. level	Correct		Incorrect		Percentages					Youden Index
	Non-event		Non-event		Correct	Sensitivity	Specificity	False-POS	False-NEG	
0.50	751	1,615	124	158	89.4	82.6	92.9	14.2	8.9	0.755
0.55	730	1,639	100	179	89.5	80.3	94.2	12.0	9.8	0.745
0.60	701	1,654	85	208	88.9	77.1	95.1	10.8	11.2	0.722
0.65	675	1,668	71	234	88.5	74.3	95.9	9.5	12.3	0.702
0.70	648	1,684	55	261	88.1	71.3	96.8	7.8	13.4	0.681
0.75	618	1,696	43	291	87.4	68.0	97.5	6.5	14.6	0.655
0.80	587	1,705	34	322	86.6	64.6	98.0	5.5	15.9	0.626
0.85	539	1,716	23	370	85.2	59.3	98.7	4.1	17.7	0.580
0.90	483	1,727	12	426	83.5	53.1	<u>99.3</u>	2.4	19.8	0.524
0.95	382	1,735	4	527	79.9	42.0	99.8	1.0	23.3	0.418

Note. The bolded cell entries (maximum values) under the “Youden Index” column indicate the values used to isolate the two single probabilities (underlined cell under the “probability level” column) that were used to define the “predicted status” according to the “Optimal Probability” rule assigning method, the method specified in section B of Table 3. Similarly, the cells with bolded and underlined entries under the percentage “correct” and percentage “specificity” columns indicate the values used to isolate the two single probabilities (for an Accept event and for a Reject event) that were used to define the “predicted status” according to the “Maximum Specificity” rule assigning method, described above and specified in section C of Table 3.

Figure 1. ROC curve plots for Mathematics and Reading item decisions.

(maximum values) under the “Youden Index” column indicate the values used to isolate the two single probabilities (underlined cell under the “probability level” column) that were used to define the “predicted status” according to the “Optimal Probability” rule assigning method, the method specified in section B of Table 3. Similarly, the cells with bolded and underlined entries under the percentage “correct” and percentage “specificity” columns indicate the values used to isolate the two single probabilities (for an Accept event and for a Reject event) that were used to define the “predicted status” according the “Maximum Specificity” rule assigning method, described above and specified in section C of Table 3.

Figure 1 provides a graphic display of how the isolated probabilities from each of the three methods fell on their respective ROC curves for mathematics decisions (top

two panels). For comparison, the corresponding results from the reading analyses are included in Figure 1 (bottom two panels). The horizontal axes (1 – specificity) for these curves, represents false positives. True positives (sensitivity) are represented on the vertical axes. All plots show that the base method and the “optimal” method result in higher levels of true positives than the “maximum specificity” method. However, the plots also show that this result comes at the price of higher levels of false positives.

Classification Accuracy and Classification Errors

Classification “accuracy” here refers to agreement between the observed status (Accept, Reject, return to FT) assigned by the analysts and the status “predicted” as a result of applying the specific method. For the training and evaluation datasets of mathematics and reading items across the first two forms of decision rules, accuracy was consistently found in the 83% to 87% range. The largest differences in accuracy between the training and evaluation datasets ranged from 0.7% (Mathematics-Base Probabilities Method) to 1.7% (Mathematics-Optimal Probabilities Method). A parallel comparison of accuracy between the training set and the evaluation set could not be completed for the third method of selecting a prediction status since that method introduced two additional predicted status categories—“Reject with review” and “Accept with review”—that were not in the options available to human reviewers.

Classification “errors,” that is, disagreements between the observed status assigned by analysts and predicted status assigned by models using the “Base Probabilities” method and the “Optimal Probabilities” method, are shown in the underlined cells of Table 6. Cells with bolded and underlined entries contain severe disagreements (e.g., accept vs. reject). The most severe of these, referred to here as Type I (model Accepts and HR Rejects) are in the upper right cells of each section of the table. Accepting these “errors” without further review would run the risk of adding poor items to a pool. The corresponding severe errors, referred to here as Type II, are those in which the model Rejects and HR Accepts (lower left cell of each section of Table 6). Accepting these “errors” without further review would run the risk of eliminating acceptable items from a pool—a waste of resources. Comparing the errors of the two types of rules reveals that the “Optimal Probabilities” rules resulted in lower levels of Type I errors but higher levels of Type II errors than the “Base Probabilities” rules. However, while these error rates would appear to be acceptably low for many prediction applications, they were deemed to be unacceptably high for the present application.

Application of the “Maximum Specificity” rules with hedge categories had the effect of reducing the levels of severe (Type I and Type II) errors. These differences can be seen by comparing the “Accept” cells and the “Reject” cells in Table 7 with the corresponding cells in Table 6. Obviously, these reductions were realized by invoking the rules created to form the “with Review” hedge categories. The model

Table 6. Confusion Matrices for the Mathematics Training Set Items and Evaluation Set Items When Predictions Are Made Using the Base Probabilities Decision Rules and the “Optimal” Probabilities Decision Rules.

HR assigned	Training set				Evaluation set			
	Predicted HR category				Predicted HR category			
	Frequency	percent	row	pct	Reject	Return to FT	Accept	Total
Base probabilities decision rules								
Reject					720	71	118	909
					27.2	2.7	4.5	34.3
					79.2	7.8	<u>13.0</u>	76.3
Return to FT					47	36	70	153
					<u>1.8</u>	1.4	<u>2.6</u>	5.8
					30.7	23.5	45.8	23.5
Accept					58	63	1465	1,586
					2.2	2.4	55.3	59.9
					<u>3.7</u>	4.0	92.4	3.8
Total					825	170	1653	2,648
					31.16	6.42	62.42	100
Optimal probabilities decision rules								
Reject					797	60	52	909
					30.1	2.3	2.0	34.3
					87.7	6.6	<u>5.7</u>	82.8
Return to FT					68	29	56	153
					<u>2.6</u>	1.1	<u>2.1</u>	5.8
					44.4	19.0	<u>36.6</u>	44.9
Accept					121	97	1368	1,586
					4.6	3.7	51.7	59.9
					<u>7.6</u>	6.1	86.3	7.0
Total					986	186	1476	2,648
					37.2	7.0	55.7	100

Note. Classification “errors,” that is, disagreements between the observed status assigned by analysts and predicted status assigned by models using the “Base Probabilities” method and the “Optimal Probabilities” method, are shown in the underlined cells of Table 6. Cells with bolded and underlined entries contain severe disagreements (e.g., accept vs. reject).

Table 7. Confusion Matrices for the Training Set Items and Evaluation Set Items When Predictions Are Made Using the Maximum Specificity Probabilities Decision Rules With Hedge Categories Included.

HR assigned	Training set							Evaluation set								
	Predicted HR category							Predicted HR category								
	Frequency	percent	row	pct	Rej	Rej/Rev	Return to FT	Acc	Acc/Rev	Total	Rej	Rej/Rev	Return to FT	Acc	Acc/Rev	Total
Reject	511	19.3			110	4.2	226	29	33	909	251	58	105	24	27	465
							8.5	1.1	1.3	34.3	17.9	4.1	7.5	1.7	1.9	33.1
		56.2			12.1		24.9	3.2	3.6		54.0	12.5	22.6	5.2	5.8	
Return to FT	9				17		76	36	15	153	8	5	51	16	18	98
							2.9	1.4	0.6	5.8	0.6	0.4	3.6	1.1	1.3	7.0
		0.3			0.6		49.7	23.5	9.8		8.2	5.1	52.0	16.3	18.4	
Accept	5.9				11.1		190	1,169	207	1,586	2	9	110	84	637	842
	7				13		7.2	44.2	7.8	59.9	0.1	0.6	7.8	6.0	45.3	59.9
		0.3			0.5		12.0	73.7	13.1		0.2	1.1	13.1	10.0	75.7	
		0.4			0.8		492	1,234	255	2,648	261	72	266	124	682	1405
Total	527				140		18.6	46.6	9.6	100	18.6	5.1	18.9	8.8	48.5	100
	19.9				5.3											

Note. Application of the “Maximum Specificity” rules with hedge categories had the effect of reducing the levels of severe (Type I and Type II) errors. These differences can be seen by comparing the “Accept” cells and the “Reject” cells in Table 7 with the corresponding cells in Table 6.

“Accept” rates and “Reject” rates in Table 7 are slightly lower in most cases, even when they are combined with their adjacent hedge category.

Conclusions and Recommendations

The core idea used in this study is not new. It is well established in the literature focusing on modeling decision-making behavior (e.g., Swets et al., 2000). This study is the first we are aware of to apply these methods to understanding how FT item performance information is considered by human analysts. This is rather uncharted, or at least undocumented, territory. At the very least, application of these methods could provide test developers with a tool to monitor HR item judgments for consistency both within and across reviewers. In the best case, these methods provide a kernel from which a virtual HR process could be developed to substantially reduce the need for actual human reviews. Such a system would be very consistent by virtue of the underlying model and would be tractable when anomalies arise.

The principal question under consideration in this study was the extent to which analysts’ judgments of FT items’ operational worthiness can be predicted from the statistical and graphical information they use to make those judgments. The most direct answer to this question is that those judgments are strongly predictable and are quite accurate across content areas, though the models required to come to those judgments vary slightly by content area, judgment type, and the combinations of available variables used. Misclassification levels were reduced by up to about 50% from the low levels obtained when using what might be considered more conventional decision-making rules. This reduction was obtained by applying decision-making rules that incorporated explicit criteria to detect predictions based on what might be termed “less indicative” sets of probabilities. A full-scale adoption of these procedures would be expected to yield substantial improvements in efficiency in the evaluation portion of what might be termed a conventional calibration review process.

Data for the study came from calibration review records. Each record was a report on the calibration results for an item and was composed of statistical data commonly associated with IRT item calibration and additional descriptive statistical information provided to analysts. An analyst’s judgment about the item (i.e., to accept it, to reject it, or to return it to FT status) was the outcome of each encounter with an item, and was the dependent variable in the effort to build models of the analysts’ judgments. For each subject area, a two-part training model was constructed, each part using stepwise logistic regression to estimate the probability that the item was (a) placed into the “Accept” status or (b) placed into the “Reject” status. Results from these training models showed the constellation of item level information provided to analysts was highly predictive of their judgments of item status. The logistic regression results also revealed that the makeup of predictors selected for each model varied slightly across subject areas and across decision types (i.e., Accept, Reject).

Table 8. Estimated Impact of Adopting the Recommended Approach.

		Decision						
Scale	Process		Reject	Rej/Rev	Return to FT	Acc/Rev	Accept	Total
Mathematics	Standard	<i>n</i>	164		948		292	1405
		%	11.7		67.5		20.8	
	Proposed	<i>n</i>	261	<u>72</u>	<u>266</u>	<u>124</u>	682	
		%	18.6	<u>5.1</u>	<u>18.9</u>	<u>8.8</u>	48.5	
	<u>Review Sum = 462 (32.9%)</u>							
	<u>Review Difference = −67.1%</u>							
Reading	Standard	<i>n</i>	166		432		192	789
		%	21.0		54.8		24.3	
	Proposed	<i>n</i>	254	<u>53</u>	<u>123</u>	<u>56</u>	303	
		%	32.2	<u>6.7</u>	<u>15.6</u>	<u>7.1</u>	38.4	
	<u>Review Sum = 232 (29.4%)</u>							
	<u>Review Difference = −70.6%</u>							

Note. Underlined cells represent the differences in the numbers of items that would be reviewed under the proposed process compared to those that would be reviewed under the standard - all items reviewed process. E.g., for mathematics (Rej w/Rev + Return to FT + Acc w/Rev) = 462 items; 32.9% of all 1405 items and 67.1% fewer items that would have to be reviewed.

To formulate the four probabilities (Accept, not-Accept, Reject, not-Reject) from the two models into a single prediction for each item, three sets of decision meta-rules were examined. The sets were developed from the estimated probabilities to form decision tables. The first set was based simply on using the probability of .5 as the cut point individually for both Accept and Reject. The second set used the optimal probabilities for each outcome as identified from the ROC stage of the analyses. The third set used the ROC analyses to find an effective balance between specificity (true negatives) and percent of correct classifications that would also minimize classification differences from the observed items status as assigned by analysts. Of these three methods, the third was considered to have the most practical advantage in that it lead to substantially lower rates of misclassifications (putative Type 1 errors).

From the results, a logical follow-up question might be, what changes to the item calibration process could be expected, if the decision meta-rules used for the third method were implemented? The answer to this question can take a number of paths, depending on the level of confidence that one is willing to invest in the outcomes shown under the evaluation set in Table 8. A very high level of confidence, for example, may suggest that the “Accept with review” and “Reject with review” categories simply be ignored and treated as “Accept” and “Reject,” respectively. That is, accept any item falling in the “Accept” or “Reject” categories as if it had been viewed by a live analyst, making further human review unnecessary. This would leave about 18% of items that would be returned to the FT cue. These items could either be reviewed by an analyst to see if an adjustment is needed for the

precalibration difficulty estimate or such adjustments could be made through a set of heuristics based on the observed data. Alternatively, the data in Table 8 could be treated with more caution. For example, items placed into the hedge categories (“Accept with review” and “Reject with review”) would always be sent to be reviewed by an analyst, while all “Accept” and all “Reject” items would be assigned these statuses without further review. Using this approach would result in roughly an additional 13% of items being reviewed. In the immediate term, the most prudent approach to the items categorized as “Return to FT” would be to continue to have them reviewed by analysts.

The latter option would appear to strike an appealing balance between accepting the models’ “judgments” and maintaining a level of human review when model probabilities indicate less certainty. In addition, this option would allow for the results of HRs to inform adjustments to the width of the hedge categories, if needed. The estimated impact of implementing this option is illustrated in shaded area of Table 8.

The estimated *ns* for the “standard” process in each content area in Table 8 were computed from the percentages observed in each decision category from the complete set of 11,811 mathematics item reviews and 4,977 reading item reviews available for the study. This provides a rough estimate of how items are likely to be assigned by analysts using the standard review process. Corresponding values for the “proposed” process were taken directly from the “total” rows from the Evaluation set in Table 7. Examination of Table 8 reveals that the straight-out Reject or Accept decisions would increase from 32% to 65% in mathematics and from 45% to 71% in reading. These increases would, of course, be accompanied by equivalent decreases in the number of items that would trigger an HR, even assuming that items judged as “returned to FT” would require a human review. If effective heuristics could be developed to reset precalibration item difficulty estimates to better target FT items to examinees, the percentage of items requiring HR would be estimated to fall into the 12% to 15% range.

Future investigation in this area would benefit from the pursuit of several other intriguing questions that would extend or validate the performance of the predictive models. For example, would different analysts consider the same item report information in the same manner? Could the procedures of the pretest item calibration process that are embedded in an operational CAT be adjusted to provide less noisy FT data in the first place? A preliminary examination by He, McCall, Thum, and Hauser (2011) suggests that this is plausible. What different forms of information could be added to an item FT report that would improve model performance? What methods can be developed to extract statistical information from graphical plots that quantifies visual information in the plots themselves that is distinct from the statistical information used to create the plots? How can such information be of service to the item evaluation process? How small can training datasets be and still yield dependable results? Is there a “drift” over time in how the information is being employed? Presumably changes in the decision-making behavior of the analyst can be expected with increased experience. This, and other changes in the analyst behaviors, will

need to be investigated in future research and reviews in order to mitigate any negative impact on the accuracy and validity of the processes used in this study.

Authors' Note

The views expressed are solely those of the authors and do not necessarily reflect the positions of NWEA.

Acknowledgment

The authors wish to thank NWEA for their support of this research.

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) received no financial support for the research, authorship, and/or publication of this article.

References

- Armstrong, J. S. (2001). Judgmental bootstrapping: Inferring experts rules for forecasting. In J. S. Armstrong (Ed.), *Principles of forecasting: A handbook for researchers and practitioners*. Norwell, MA: Kluwer Academic.
- Dawes, R. M. (1979). The robust beauty of improper linear models in decision making. *American Psychologist*, 34, 571–582.
- Dawes, R. M. (1993). Prediction of the future versus understanding of the past: A basic asymmetry. *The American Journal of Psychology*, 106 (1), 1–24.
- Dawes, R. M., Faust, D., & Meehl, P. E. (1989). Clinical versus actuarial judgment. *Science*, 243, 1668–1674.
- Drasgow, F., Levine, M. V., Tsien, S., Williams, B., & Mead, A. D. (1995). Fitting polytomous item response theory models to multiple-choice tests. *Applied Psychological Measurement*, 19, 143–165.
- Einhorn, H. J. (1974). Expert judgment: Some necessary conditions and an example. *Journal of Applied Psychology*, 59, 562–571.
- Glas, C. A. W., & Suarez-Falcon, J. C. (2003). A comparison of item fit statistics for the three-parameter logistic model. *Applied Psychological Measurement*, 27, 87–106.
- Goldberg, L. R. (1968). Simple models or simple processes? Some research on clinical judgments. *American Psychologist*, 23, 482–496.
- Hambleton, R. K., & Han, N. (2005). Assessing the fit of IRT models to educational and psychological test data: A five step plan and several graphical displays. In W. R. Lenderking & D. Revicki (Eds.), *Advances in health outcomes research methods, measurement, statistical analysis, and clinical applications* (pp. 57–78). Washington, DC: Degnon Associates.

- Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of item response theory*. Newbury Park, CA: Sage.
- He, W., McCall, M., Thum, Y. M., & Hauser, C. (2011, April). *Pretest item calibration for computerized adaptive tests*. Paper presented at the annual meeting of the National Council on Measurement in Education, New Orleans, LA.
- Libby, R. (1976). Man versus model of man: The need for a non-linear model. *Organizational Behavior and Human Performance*, 16, 1–12.
- Meehl, P. E. (1954). *Clinical versus statistical prediction: A theoretical analysis and a review of the evidence*. Minneapolis: University of Minnesota Press.
- Orlando, M., & Thissen, D. (2000). Likelihood-based item fit indices for dichotomous item response theory models. *Applied Psychological Measurement*, 24, 50–64.
- Orlando, M., & Thissen, D. (2003). Further investigation of the performance of $S - X^2$: An item fit index for use with dichotomous item response theory models. *Applied Psychological Measurement*, 27, 289–298.
- Schisterman, E. F., Perkins, N. J., Liu, A., & Bondell, H. (2005). Optimal cut-point and its corresponding Youden index to discriminate individuals using pooled blood samples. *Epidemiology*, 16, 73–81.
- Shtatland, E. S., Kleinman, K., & Cain, E. M. (2004, May). *A new strategy of model building in Proc Logistic with automatic variable selection, validation, shrinkage and model averaging*. Paper presented at the SAS Users Group International Conference, Montreal, Quebec, Canada.
- Smith, R. M. (2000). Fit analysis in latent trait models. *Journal of Applied Measurement*, 1(2), 199–218.
- Stone, C. A., & Zhang, B. (2003). Assessing goodness of fit of item response theory models: A comparison of traditional and alternative procedures. *Journal of Educational Measurement*, 40, 331–352.
- Swets, J. A., Dawes, R. M., & Monahan, J. (2000, October). Better decisions through science. *Scientific American*, 283, 82–87.
- Weiss, D. J., Shanteau, J., & Harries, P. (2006). People who judge people. *Journal of Behavioral Decision Making*, 19, 441–454.