DOCUMENT RESUME

ED 392 797 TM 023 782

AUTHOR DeMauro, Gerald E.
TITLE Construct Validation of Minimum Competence in Standard Setting. Revised.
PUB DATE May 95
NOTE 29p.; Paper presented at the Annual Meeting of the National Council on Measurement in Education (San Francisco, CA, April 19-21, 1995).
PUB TYPE Reports - Research/Technical (143) -- Speeches/Conference Papers (150)

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS Ability; Competence; *Construct Validity; Difficulty Level; High Schools; *High School Students; Interrater Reliability; *Judges; *Minimum Competency Testing; Multiple Choice Tests; Scores; Standards; *Test Items
IDENTIFIERS *Angoff Methods; New Jersey; Ninth Grade Proficiency Test; *Standard Setting

ABSTRACT

Studies of the Angoff method of standard setting suggest that judges agree in their estimates of the relative difficulties of test questions for minimally competent examinees and that each judge's estimates correlate well with the observed item difficulties for examinees whose total test scores are near the judge's personal standard (G. E. DeMauro, 1991). This finding suggests that Angoff estimates contain additive item-related and judge-related components, varying both from judge to judge and from estimated to observed performance by constants. Since, in homogeneous tests, observed performance on items also varies by constants over ability levels, the observed convergence of each judge's estimates on item performance near an individual standard is really a special case of convergence of all judges on item performance near a common deliberated standard. Data from the New Jersey High School Proficiency Test (NJHSPT) standard setting study supported this hypothesis. The convergence of the judges on a construct of minimal competence was studied for the standard setting study of multiple-choice items for three tests of the NJHSPT for grade 11. In all, 78 judges were involved. (Contains 3 tables and 14 references.) (Author/SLD)

Construct Validation of Minimum Competence in Standard Setting[1]

Gerald E. DeMauro
New Jersey State
Department of Education

Paper presented at the annual meeting of the National Council of Measurement in Education, San Francisco, April, 1995.

Revised, May 1995

[1]I wish to acknowledge the valuable suggestions of Dr. Jason Millman, based on his review of an earlier draft.

Abstract

Studies of the Angoff method of standard setting suggest that judges agree in their estimates of the relative difficulties of test questions for minimally competent examinees, and that each judge's estimates correlate well with the observed item difficulties for examinees whose total test scores are near the judge's personal standard (DeMauro, 1991). This finding suggests that Angoff estimates contain additive item-related and judge-related components, varying both from judge to judge and from estimated to observed performance by constants. Since, in homogenous tests, observed performance on items also varies by constants over ability levels, the observed convergence of each judge's estimates on item performance near an individual standard is really a special case of convergence of all judges on item performance near a common deliberated standard. Data from the New Jersey High School Proficiency Test standard setting study support this hypothesis.

Construct Validation of Minimum Competence in Standard Setting

Gerald E. DeMauro
New Jersey State Department
of Education

Introduction

The Angoff (1971) procedure produces reasonable standards based on the ability of judges to first define a hypothetical population of minimally competent examinees and then to estimate item difficulty for this group. While there have been extensive evaluations of the estimates (DeMauro and Powers, 1990; DeMauro, 1991) there have been few systematic studies of the definition of the construct of the hypothetical examination group. This study attempts to apply the methodology of the evaluative studies to the issue of validating the construct of minimal competence.

This construct is complex in the sense that it is defined not only by the content domain of the test, but also by a hypothetical point in the distribution of skills of the examinees. Therefore, construct validation must address the capacity of judges to converge on the minimum competency construct both in terms of the content domain to be sampled and the level of skills needed to demonstrate competence. As the Angoff process involves increasing elaboration of the construct, there should be evidence of increasing convergence through the course of the judges' deliberations.

There is a growing body of evidence supporting the ability of judges to estimate the relative difficulty of test items, even though they may disagree about the absolute difficulty of those items (Brennan & Lockwood, 1980; Skakum & Kling, 1980; DeMauro, 1991). This implies that there is some component of each estimate that is item-based that is reliably discerned across judges and another component that is judge-based. If there exists some true ability measure of the hypothetical group of minimally competent examinees, it would mean that judges are not idiosyncratic in how they perceive this group, but share some common understanding of the skills of this group. Put more empirically, we would expect that differences in the estimates of item difficulties for this hypothetical group of examinees are not related to the items themselves, but are related to the judges, and that there are no reliable judge by item interactions, e.g., the judge and item effects are additive, in the estimated item difficulties required by the Angoff procedure.

Angoff judgments, then, can be evaluated via a repeated measures analysis, in which there are main effects for item (repeated measure) and for estimation variability of judges and an interactive error effect for judge by item. To borrow from

reliability theory (Kerlinger, 1964), the interaction term should be much smaller than either of the main effects, and reliability across judges or reliability of judgments across items should be the difference of the mean squares for the main effect and the interaction divided by the mean square for the main effect.

Estimated and Observed Item Difficulties

This view of reliability is actually a special case of Kane's (1986) notion that a group can be chosen about each judge's estimated passing standard, and the observed item difficulties for this group can be correlated with the judges' Angoff estimates as an evaluation of reliability of judgments.

A second analysis, then can be made of the correlation between estimated and observed item difficulties. A high correlation would indicate that for each judge, the difference between the two approaches some constant, k, which can be thought of as an estimating effect. Any interaction in estimated difficulties between this judge estimation variable and test items would decrease the magnitude of observed and estimated item difficulties.

Specifically, Kane (1986) and DeMauro and Powers (1990) propose that each judge's Angoff estimates should agree with the observed item difficulties for actual examinees near that judge's standard. A homogenous test, that is, one that measures the same construct throughout the range of scores, would have item difficulties that are highly correlated at various score intervals (Angoff & Modu, 1973). Therefore, the correlation of each judge's estimated item difficulties with observed difficulties at that individual judge's estimated passing standard should approximate the correlation of each judge's estimated difficulties with the observed difficulties at the average, deliberated passing standard computed across judges.

Naturally, this is an ideal that holds for a completely homogenous test, and is observed in varying degrees for each panel of judges, for each test under consideration. However, it does speak to the very basic assumption of the Angoff methodology, that there are some true item difficulties for the hypothetical group of minimally competent examinees, and the interaction of items and the estimating deviations associated with individual judges must be small enough to permit item difficulty estimates to be averaged across judges to obtain the overall passing standard. The veracity of this assertion is borne out by the observations that judges agree in their estimates of the relative difficulty of items; that is, any item by judge interactions are so small as to not alter the difficulty ordering of the items.

Elaboration of the Construct

As the judges converge on the construct of minimal competence in the test domain, the correlation between estimated and observed item difficulties should also improve. This assertion follows from the reliability argument. For example, both estimated and observed item difficulties share a true item difficulty component (largely comprising the item effect described above), the judge estimation error described above, an item by judge interaction effect (hypothesized to be small), and a construct elaboration effect and its interactions with judge, item, and judge by item components.

The construct elaboration effect could be conceived as changes in estimations associated with greater elaboration of the construct. As judges come to understand a shared, deliberated view of minimal competence, the agreement of estimated and observed item difficulties should increase because of reduction of this variable.

As judges converge on the construct, we would expect this variable and all possible interactions with it to approach zero. Hence, judges' estimates would approach observed difficulties plus some estimation constant associated with each judge. This constant is the value, k, described above for observed difficulties sampled near the judge's individual standard. Convergence on the construct of minimal competence, which is implied by the observed agreement of relative difficulty of test items, also implies that there is another constant, k', which reliably describes the difference between the item performance of examinees close to the average, deliberated standard across judges and the estimated difficulty for each judge.

On the other hand, if an interaction was observed between judge and item, then estimated difficulties would not vary from observed difficulties for each judge by the constants described above, and the item and judge reliabilities would drop and the correlation between observed and estimated item difficulties both around the standard derived by averaging across judges and the standard derived from each individual judge's estimates would also drop because the shared item difficulty component of observed and estimated item difficulties would be relatively smaller. The judge's estimates would sometimes overestimate and sometimes underestimate. Where there were underestimates, the judge would be estimating the difficulty for a poorer skilled group than minimally competent examinees. Where there were over estimates, the judge would be estimating the difficulty for a better-skilled group than minimally competent. Hence, the reliability of the judges' estimates and their correlations with observed item difficulties speak directly to the construct of minimal competence and the construct validity of the procedure.

Proposed Model of Judgments

We must remember that the Angoff method requires judgments to be made at the item level. Convergence on the construct, then, would require convergence at the item level, with estimation errors associated with judges distributed about the mean estimated item difficulties. This is why it is reasonable to expect that the effect for judges could be measured about the standard set across judges, and not just about each individual judge's standard. Simply put, the model we are proposing first hypothesizes that each judge's estimates are based on an increasingly shared construct of minimal competence, not an individually-defined construct. Differences in estimated difficulties are differences associated with each judge's capacity to estimate accurately the item by item performance of the hypothetical group of minimally competent examinees. Because we do not expect these estimation errors to interact with test items, the estimated difficulties for a hypothetical group of minimally competent examinees for each judge should correlate at least as well with observed difficulties for examinees close to the standard based on the overall deliberated standard, averaged across judges, as they do for observed difficulties for examinees close to each judge's individual standard.

This formulation may be empirically represented as :

$$\Delta_e = \Delta_a + \Delta_{k'} + \Delta_i + \Delta_c + \Delta_{k'c} + \Delta_{ic}$$

where Δ_e = Angoff estimated item difficulties

Δ_a = Actual item difficulty of hypothetical group

$\Delta_{k'}$ = Judge's estimation variable

Δ_i = Item variable

Δ_c = Construct elaboration variable

Note: We hypothesize no interactions for judge by item nor for judge by item by construct elaboration.

Focus of Inquiry

The convergence of the judges on a construct of minimal competence was studied for the standard setting study of the multiple choice items for the three tests (Reading, Mathematics, and Writing) of the New Jersey eleventh grade High School Proficiency Test (HSPT11). Judges were evaluated according to on the reliability of their Angoff estimates of item difficulties and on the agreement of these ratings with observed item difficulties.

Study Questions

Specifically, the construct was delimited with reference to the following:

1. Are the Reading, Mathematics, and Writing sections of the HSPT11 sufficiently homogenous throughout the ranges of individual judges' estimates of the standards to support the hypothesis that the same construct is is being measured within these ranges and the hypothesis that the judges were responding to the same construct in making their estimates?

2. Was there evidence that judges were reliably estimating item difficulty in terms of the relative size of the main effects associated with judges and with items relative to the interaction of judges and items?

3. Were the relative sizes of these effects consistent for both initial (less elaborated construct) Angoff estimates and for the final (fully elaborated construct) estimates?

4. Is the correlation of initial item difficulty estimates for the hypothetical group of examinees with each judge higher with observed performance of examinees near the ultimate averaged, deliberated standard across judges than it is with either observed performance for examinees nearest each judge's individual standard or observed performance for examinees nearest the initial averaged standard?

5. Is the agreement of final item difficulty estimates greater with the ultimate deliberated standard than it is with the final individual standard for each judge?

METHODS

Classification of Judges and Observed Performance Sampling

To simplify the analyses, as well as to provide some stability of the statistics, observed item difficulties were computed based on population quintiles for the whole raw test score. This is done in hopes that what is gained in the greater reliability for observed item difficulties (which here serve as a proxy for true difficulties for the hypothetical examinee group) compensates for the loss of accuracy in the expanded sampling over the quintile. The correlations between observed and estimated item difficulties may be depressed somewhat by this sampling of examinees, although the extent of this depression should be slight for a homogenous test.

All item difficulty values for the analyses were expressed on the delta scale (Angoff & Modu, 1973) to provide equal interval scale properties. The delta scale is a normal transformation of p-values with a mean of 0 and a standard deviation of -4.

The Instrument

The standard setting study that provides the data for these analyses was conducted on the New Jersey eleventh grade High School Proficiency Test (HSPT11), from December 15 to December 17, 1993. This study explicitly addresses the Angoff procedure used to estimate standards for the multiple choice test questions. Other studies (Webb & Miller, 1995) address the procedures used with the open-ended questions.

The instrument consists of three tests: Reading, Mathematics, and Writing. On the form of the instrument used for the standard setting study, there were 37 Reading items, 31 Mathematics items, and 36 Writing items that all focused on Revising and Editing skills.

Each has both open-ended and multiple choice components, and the Mathematics test also has two grid-response items that are treated with the multiple choice items in this analysis. Students must pass all three sections of the instrument to be eligible for graduation unless they are either exempt in relation to a Special Education classification, or demonstrate their skills through an alternative route called the Special Review Assessment (available to senior students who are otherwise eligible for graduation and who have the desired levels of skill).

The Judges

A separate panel of judges was configured for each of the three test sections. Angoff procedure judges included teachers recommended for their expertise in reading, mathematics, and writing, respectively, representatives of the business community (1-2 on each panel), two students each, and two parents each. The differences among these types of judges are not the major focus of this study, and may be reviewed in the report of the findings (New Jersey Department of Education, 1994).

In all, there were 21 judges on the reading panel, 29 judges on the mathematics panel, and 28 judges on the writing panel.

Standard Setting Procedures

The standard setting occurred over a series of phases. In the first phase, judges were asked to define the group of minimally competent examinees and to study the test. This phase includes deliberated specification of the attributes of minimally competent students, as well as actually taking the test, discussing the answers, and making preliminary Angoff estimates of the percentage of minimally competent examinees who would pass each item.

This first phase is treated in the current study as a partially elaborated construct phase. The lack of elaboration variable is thus larger than it would be for the final phase of the standard setting study. The partial elaboration phase is followed by greater elaboration which includes providing judges with item level performance data for each quintile of examinees, and negotiation among the judges concerning their estimates. Other phases include discussion about outlier judges and items. In this study, the final estimates of judges are compared to the preliminary estimates to evaluate the effect of construct elaboration.

Describing Minimal Competence

To insure that the judges all hypothesized the construct of minimal competence with respect to the HSPT11 domains, evidence is offered that their descriptions of minimal competence were linked to this domain. The delineation of the attributes of minimal competence was the first task in standard setting.

As part of the development of the Special Review Assessment process in New Jersey the standing HSPT11 content committees of experts in reading, mathematics, and writing were convened to discuss how alternative assessments might be designed to identify students who exhibited the attributes of minimal competence in

each of these areas. The first step in this process was to invite the committee members to rate the relevance of each of the attributes of minimal competence listed by the standard setting committees in each area for its relevance to the domain of the appropriate HSPT11 content area. These ratings may be viewed as the first step in the construct validation of minimum competence, to the extent that they indicate whether the attributes of the hypothetical group identified by each standard setting committee were consistent with the content domain.

The ratings were made on a five point scale in which 0 indicated not relevant, 1 indicated somewhat relevant, 2 indicated relevant, 3 indicated very relevant, and 4 indicated critical. The criterion for validating that each of the attributes belong to the content domain was an average rating of 2.5 (the upper bound of "relevant") for one or more of the content clusters measured by the HSPT11. There are four content clusters in Reading, five in Mathematics, and two in Writing. The panelists, as members of the content committees, were thoroughly familiar with these content domains and were the most knowledgeable individuals to make these linkages.

The demonstration of relevance was chosen because the joint standards (AERA, APA, & NCME, 1985) require that "relevance" of the universe represented to the use of the test must be described as a requisite of presenting content-related evidence in support of validity. In all, 23 members of the reading and the writing committees rated the reading and the writing attributes for relevance to the respective test specifications (both committees were used for each set of attributes). Six members of th mathematics committee rated the mathematics attributes for relevance to the mathematics content clusters of the HSPT11.

There were 12 reading, 7 writing, and 26 mathematics attributes. Each was linked by average ratings of 2.5 or higher to at least one or more content clusters in the respective content areas.

Homogeneity of Measure: Question 1

Item difficulties for examinees in each quintile of overall raw score were correlated. To linearize the measure of item difficulties, p-values were first converted into delta scale values (Angoff & Modu, 1973).

Reliabilities: Questions 2 and 3

A repeated measure analysis of variance was made of judges by items for both the initial and for the final estimated item difficulties. The estimates were converted to the delta scale to insure the proper interval properties. Reliability for

judges and for items was computed by subtracting the mean squares of the judge by item interaction terms from the mean squares associated with the main effects and dividing by the appropriate main effects.

A second analysis was performed, involving the construct elaboration effect. This effect was estimated by differences between the initial and final item difficulty estimates (repeated measure), the judge estimation effect, the effect, and the first and second order interactions of these three variables. Reliability is estimated for judges and items under the hypothesis that there would be no judge by item interaction and no judge by item by elaboration interaction. the mean squares for both of these interactions were summed and subtracted from the main judge and item effects and divided by the main judge by item effects.

Correlations: Questions 4 and 5

The Angoff task requires that judges estimate the percentages of the hypothetical group of minimally competent examinees that would answer each question correctly (Livingston & Zieky, 1982). This was modified in the New Jersey study by restricting these estimates to 20 options, one for each 5 percent.

To assess the agreement of each judge's estimates with the observed item difficulties, several correlations were made. First, a standard both for initial, partially-elaborated estimates and for final, fully-elaborated estimated was determined by averaging the estimated difficulties in terms of delta values for each judge. This average was associated with the nearest quintile in terms of averaged observed delta values. Missing data from initial estimates (three items across all judges and test sections) were replaced with the average delta value for that item for the initial estimates.

Overall averaged or deliberated standards were then computed for each of the three content areas by averaging the sums of the p-value ratings for each of the judges. These average standards were the actual Angoff passing standards adopted for the multiple choice sections of each test. These standards were then also located in one of the five quintiles. Note, the same quintile pertained whether it was identified in terms of the nearest average estimated delta value or the location of the averaged deliberated standards in the score distributions. These standards are called deliberated because they involve deliberation, as prescribed by the Angoff procedure, through multiple stages.

P-values were computed for students scoring within each of

the five quintiles based on the actual October 1993 administration of the HSPT11. Approximately 12,000 students each comprised each quintile. These p-values were converted to delta values, as well.

Each judge's initial and final estimates of the percentage of the hypothetical group that would answer each question correctly and the delta values for each quintile were correlated. The initial judgments were based on partial elaboration of the construct and the final judgments were based on a fuller elaboration of the construct.

Results

Homogeneity of Measure: Question 1

For Reading, judge's individual standards were located either within the first or second quintile of performance. The correlation of observed item delta values for groups in these quintiles was .976.

For Mathematics, judge's individual standards ranged from within the first to within the third quintile. The correlation of observed item deltas for groups in these quintiles were .903 for the first and second quintile, .824 for the first and third quintile, and .974 for the second and third quintile. Because the overall standard averaged over judges was in the second quintile, it is the correlations of item difficulties involving this quintile that are most germane to the current evaluation.

Finally, for Writing, the judge's standards were located either within the first or second quintile. The correlation of observed item difficulties for examinees in the first quintile with those for examinees in the second quintile was .975.

Reliabilities: Questions 2 and 3

Table 1 presents the components of variance for the repeated measure reliability analyses. As shown, the interjudge and interitem Reliabilities of the Angoff estimates were high both for the initial and final judgments. The analyses involving both initial and final estimates as a repeated measure also demonstrated high interjudge and interitem Reliabilities, even after the judge by item and judge by item by construct elaboration terms are combined. The model is supported.

TABLE 1

Components of Variance
for Reliability Evaluation of
Angoff Difficulty Estimates (ctd.)

Area	Statistic	Components		
		Judge	Item	Judge*Item
Initial Judgments				
Reading	ss	566.48	765.48	1896.54
	df	20	36	720
	ms	28.32	21.26	2.63
	reliability	.907	.876	
Mathematics	ss	997.81	1513.14	2416.46
	df	28	30	840
	ms	35.64	50.44	2.88
	reliability	.919	.943	
Writing	ss	947.02	737.27	2474.81
	df	27	35	945
	ms	35.07	21.06	2.62
	reliability	.925	.876	
Final Judgments				
Reading	ss	262.34	1113.40	437.50
	df	20	36	720
	ms	13.12	30.93	0.61
	reliability	.954	.980	
Mathematics	ss	618.61	2420.12	457.85
	df	28	30	840
	ms	22.09	80.67	0.55
	reliability	.975	.993	
Writing	ss	431.35	1448.22	891.69
	df	27	35	945
	ms	15.98	41.38	0.94
	reliability	.941	.977	

TABLE 1

Components of Variance
for Reliability Evaluation
of Angoff Difficulty Estimations

Area	Statistic	Components: Judge	Item	Judge*Item	J.*I.*Elab.
Initial Judgments					
Reading	ss	429.79	1535.02	1151.71	1182.32
	df	20	36	720	720
	ms	21.49	42.64	1.60	1.64
	reliability	.849	.924		
Mathematics	ss	1274.38	3530.19	1874.99	999.31
	df	28	30	840	840
	ms	117.63	50.44	2.23	1.19
	reliability	.925	.971		
Writing	ss	982.27	1960.44	1962.02	1404.49
	df	27	35	945	945
	ms	36.38	56.01	2.08	1.49
	reliability	.902	.936		

Question 4

Table 2 shows that, in the first stages of estimation, the judges showed some agreement with observed item difficulties, even though that had not, at that point reviewed actual examinee performance. Note that there is even convergence on the construct of minimal competence. For example, in Reading and in Mathematics, the ultimate passing standard averaged over judges was nearest the second quintile in the overall score distributions. In fact, over the 21 Reading judges, 16 had initial standards nearest to the first quintile, and of these 16, 13 had item difficulty estimates that correlated higher with observed difficulties around the final standard (the second quintile) than around than around the first quintile, which was both their own personal standard and the initial group mean standard. Of the five judges who set their initial standards nearest the mean for the second quintile, all five had item difficulty estimates that correlated highest with the observed difficulties for the second quintile.

In Mathematics, the same phenomena was observed. Of the 29 judges, 21 had initial personal standards nearest the first quintile, even though the ultimate overall standard would be set nearest the second quintile. Of these 21, 20 had item difficulties that correlated higher with observed difficulties nearest the second quintile than nearest the first. Of the eight judges whose initial item difficulty estimates were nearest the second quintile, all eight had item difficulty estimates that correlated highest with the observed difficulties for the second quintile.

For Writing, the story was somewhat different. Both the initial and the final overall standards, averaged over judges, were nearest the second quintile. Five judges had initial mean item difficulty estimates nearest the second quintile. For four of these five, however, the estimates correlated higher with observed difficulties for the first quintile than for the second. In all three areas, there was greater convergence on the point of the distribution where the ultimate standard would be set across judges than there was on either the point where their initial personal standards were located or the point in which the initial average standard, across judges, was located.

TABLE 2

Correlations between
First Item Difficulty
Estimates (in Deltas), and
Final Item Difficulty Estimates
(in Deltas), with Observed
Item Difficulties (Also in Deltas)
for Examinees Near the Average
Passing Standard, for Each Judge

Area	Judge	Item Difficulty	Nearest Quintile	Correlations of First Estimated Difficulties And Observed Difficulties for Individual Est.	1st Avg. Est.	Final Avg. Est.
Reading	a	13.02	1	.123	.123	.207
	b	12.23	1	.207	.207	.286
	c	11.24	2	.270	.173	.270
	d	12.58	1	.033	.033	.184
	e	11.60	2	.266	.184	.266
	f	12.59	1	.244	.244	.290
	g	12.69	1	-.011	-.011	.158
	h	13.97	1	.220	.220	.273
	i	13.57	1	.136	.136	.170
	j	12.98	1	-.042	-.042	.091
	k	12.85	1	.272	.272	.233
	l	11.36	2	.071	-.054	.071
	m	13.77	1	.169	.169	.068
	n	13.83	1	.317	.317	.325
	o	11.79	2	.007	.007	.026
	p	12.92	1	.334	.334	.356
	q	13.00	1	.301	.301	.336
	r	13.85	1	.304	.304	.414
	s	13.28	1	.229	.229	.222
	t	13.88	1	.088	.088	.189
	u	11.66	2	.093	.037	.093

TABLE 2

Correlations between
First Item Difficulty
Estimates (in Deltas), and
Final Item Difficulty Estimates
(in Deltas), with Observed
Item Difficulties (Also in Deltas)
for Examinees Near the Average
Passing Standard, for Each Judge

Area	Judge	Item Difficulty	Nearest Quintile	Correlations of First Estimated Difficulties And Observed Difficulties for Individual Est.	1st Avg. Est.	Final Avg. Est.
Math.	a	13.57	1	.251	.251	.305
	b	13.69	1	.525	.525	.604
	c	14.47	1	.392	.392	.407
	d	15.38	1	.209	.209	.303
	e	14.58	1	.514	.514	.657
	f	14.48	1	.450	.450	.534
	g	13.92	1	.146	.146	.234
	h	13.56	1	.493	.493	.550
	i	13.34	1	.451	.451	.458
	j	12.95	1	.411	.411	.483
	k	13.26	1	.469	.469	.564
	l	13.64	1	.459	.459	.617
	m	12.85	1	.347	.347	.476
	n	14.17	1	.170	.170	.300
	o	12.71	1	.365	.365	.372
	p	12.41	2	.421	.308	.421
	q	12.52	2	.188	.172	.188
	r	11.37	2	.783	.628	.783
	s	15.98	1	.407	.407	.419
	t	12.76	1	.279	.279	.448
	u	12.47	2	.374	.344	.374
	v	13.37	1	.345	.345	.306
	w	14.55	1	.281	.281	.440
	x	12.13	2	.275	.146	.275
	y	12.18	2	.497	.407	.497
	z	12.56	2	.481	.379	.481
	a2	13.05	1	.686	.686	.735
	b2	11.51	2	.025	-.148	.025
	c2	13.30	1	.507	.507	.562

TABLE 2

Correlations between
First Item Difficulty
Estimates (in Deltas), and
Final Item Difficulty Estimates
(in Deltas), with Observed
Item Difficulties (Also in Deltas)
for Examinees Near the Average
Passing Standard, for Each Judge

Area	Judge	Estimated Item Difficulty	Nearest Quintile	Correlations of First Estimated Difficulties And Observed Difficulties for Individual Est.	1st Avg. Est.	Final Avg. Est.
Write	a	11.72	2	.151	.202	.202
	b	12.82	1	.452	.452	.452
	c	12.52	1	.219	.219	.219
	d	11.72	2	.493	.504	.504
	e	13.01	1	.567	.567	.567
	f	14.86	1	.129	.129	.129
	g	11.51	2	.102	.051	.051
	h	13.06	1	.573	.573	.573
	i	12.97	1	.197	.197	.197
	j	12.73	1	.493	.493	.493
	k	12.60	1	.146	.146	.146
	l	14.18	1	.579	.579	.579
	m	12.67	1	.393	.393	.393
	n	13.51	1	.382	.382	.382
	o	14.67	1	.300	.300	.300
	p	13.40	1	.198	.198	.198
	q	12.26	1	.559	.559	.559
	r	11.29	2	.336	.365	.365
	s	13.41	1	.527	.527	.527
	t	13.02	1	.167	.167	.167
	u	13.03	1	.381	.381	.381
	v	14.62	1	.496	.496	.496
	w	11.07	2	.190	.223	.223
	x	13.10	1	-.100	-.100	-.100
	y	12.25	1	.254	.254	.254
	z	13.11	1	.343	.343	.343
	a2	14.14	1	.338	.338	.338
	b2	12.24	1	.425	.425	.425
	c2	13.30	1	.507	.507	.507

Question 5

As is shown in Table 3, for most judges in all three test sections, their final individual standard was nearest the same quintile as the average standard. Therefore, the correlations between estimated and observed item difficulties for examinees in the quintile nearest the judge's estimates and the quintile nearest the averaged deliberated estimate was the same.

In Reading, eight judges had standards that were closer to other quintiles than that of the average standard. Among these eight judges, five had item difficulty correlations higher based on the average standard than based on their individual standards. In Mathematics, of the ten judges whose individual standards were closer to a different quintile than the average standard, seven had item difficulties correlations higher based on the average standard than based on their individual standards. In Writing, four judges had individual standards closer to different quintiles than the average standard. Among these four, two had item difficulty correlations higher based on the average standard than based on their individual standards.

TABLE 3

Correlations between Item Difficulty Estimates (in Deltas) and Observed Item Difficulties (also in deltas) for Each Judge's Individual Passing Standard and for the General Averaged Standard

Area	Judge	Estimated Item Difficulty	Nearest Quintile	Correlations with Difficulties at Judge's Quintile	Correlations with Difficulties at Average Quintile	Judge's Standard
Reading (Avg. quint. = 2)	a	10.75	2	.940	.940	25.65
	b	10.81	2	.773	.773	25.80
	c	12.50	1	.581	.717	20.25
	d	11.75	2	.903	.903	22.85
	e	12.18	1	.860	.884	21.35
	f	12.37	1	.791	.857	20.70
	g	12.87	1	.386	.377	18.95
	h	11.73	2	.872	.872	22.75
	i	11.69	2	.861	.861	22.95
	j	12.64	1	.688	.694	19.75
	k	11.22	2	.778	.778	24.60
	l	11.84	2	.906	.866	22.55
	m	11.83	2	.872	.872	22.55
	n	11.65	2	.854	.854	23.10
	o	11.17	2	.853	.853	24.45
	p	11.67	2	.911	.911	22.95
	q	12.50	2	.586	.597	20.25
	r	12.55	1	.638	.629	20.05
	s	11.41	2	.831	.831	24.00
	t	12.02	2	.944	.944	21.95
	u	11.38	2	.831	.831	23.90
Average		11.83				22.45

TABLE 3

Correlations between Item Difficulty Estimates (in Deltas) and Observed Item Difficulties (also in deltas) for Each Judge's Individual Passing Standard and for the General Averaged Standard

Area	Judge	Estimated Item Difficulty	Nearest Quintile	Correlations with Difficulties at Judge's Quintile	Correlations with Difficulties at Average Quintile	Judge's Standard
Math. (Avg. quint. = 2)	a	13.57	1	.895	.858	13.80
	b	12.05	2	.905	.905	17.95
	c	12.58	2	.904	.904	16.70
	d	13.33	1	.783	.862	14.50
	e	13.56	1	.900	.930	13.80
	f	13.85	1	.773	.837	13.05
	g	11.49	2	.968	.968	19.65
	h	13.00	1	.878	.885	15.45
	i	12.15	2	.907	.907	17.95
	j	11.93	2	.895	.895	18.55
	k	12.62	2	.902	.902	16.60
	l	12.18	2	.932	.932	17.75
	m	12.60	2	.910	.910	16.55
	n	12.57	2	.861	.861	16.75
	o	11.33	2	.924	.924	19.95
	p	11.62	2	.847	.847	19.45
	q	12.58	2	.844	.844	16.60
	r	11.43	2	.909	.909	19.90
	s	12.87	1	.884	.948	15.70
	t	11.99	2	.896	.896	18.30

TABLE 3

Correlations between Item Difficulty Estimates (in Deltas) and Observed Item Difficulties (also in deltas) for Each Judge's Individual Passing Standard and for the General Averaged Standard

Area	Judge	Estimated Item Difficulty	Nearest Quintile	Correlations with Difficulties at Judge's Quintile	Correlations with Difficulties at Average Quintile	Judge's Standard
Math.	u	11.98	2	.964	.964	18.35
	v	10.27	3	.857	.794	23.00
	w	13.18	1	.730	.844	15.00
	x	10.79	3	.939	.921	21.40
	y	12.53	2	.878	.878	16.80
	z	11.88	2	.957	.957	18.70
	aa	12.24	2	.970	.970	17.60
	bb	11.39	2	.887	.887	20.05
	cc	13.03	1	.848	.932	15.35
Average		12.30				17.42
Writ. (Avg. quint. = 1)	a	12.60	1	.874	.874	19.20
	b	12.55	1	.723	.723	19.55
	c	12.66	1	.772	.772	19.15
	d	12.32	1	.847	.847	20.15
	e	11.43	2	.581	.606	23.15
	f	12.96	1	.722	.722	18.15
	g	12.85	1	.918	.918	18.45
	h	12.77	1	.983	.983	18.70
	i	12.80	1	.839	.839	18.60
	j	13.07	1	.703	.703	17.75
	k	12.67	1	.698	.698	19.05
	l	12.68	1	.831	.831	19.05
	m	13.06	1	.842	.842	17.80
	n	12.94	1	.924	.924	18.20
	o	11.67	2	.935	.930	22.30
	p	12.66	1	.848	.848	19.15
	q	12.23	1	.817	.817	20.65

TABLE 3

Correlations between Item Difficulty Estimates (in Deltas) and Observed Item Difficulties (also in deltas) for Each Judge's Individual Passing Standard and for the General Averaged Standard

Area	Judge	Estimated Item Difficulty	Nearest Quintile	Correlations with Difficulties at Judge's Quintile	Correlations with Difficulties at Average Quintile	Judge's Standard
Writ.	r	11.57	2	.628	.655	22.85
	s	12.24	1	.915	.915	20.45
	t	12.69	1	.831	.831	19.05
	u	12.87	1	.731	.731	18.40
	v	14.58	1	.635	.635	12.55
	w	11.48	2	.852	.821	23.05
	x	13.16	1	-.100	-.100	17.35
	y	11.48	2	.809	.770	23.00
	z	13.06	1	.653	.653	17.80
	aa	13.01	1	.946	.946	17.90
	bb	12.03	1	.675	.675	21.30
Average		12.57				19.38

Discussion

In general, it appears that the three instruments are homogenous throughout the range of judge's individual standards based on estimates of the performance of a hypothetical group of minimally competent examinees. This means that if, in fact, judges differ in their estimates by some constant, which is consistent with the literature, than they will correlate as well with the observed item difficulties near the overall standard based on the average of individual standards as they will with observed item difficulties near their own individual standards.

This hypothesis is further supported by the finding that the interaction of judges' estimates with items is very small, yielding high reliability coefficients both for the judges and for the items in each of the three tests. Moreover, the consistency of this finding, both for the initial difficulty estimates, for the final estimates, and for both estimates suggests that the judges were acting consistently in their judgments, even in the initial phases, with a construct of minimal competence.

The correlations between observed and estimated item difficulties support the hypotheses. In both the initial judgments and in the final judgments, the correlations were generally as high for the item difficulties based on the average standards as they were based on the individual standards.

Lower correlations for the initial judgments reflect not a different construct of minimal competence, but rather less inability to make consistent judgments across items. Hence, if the items were judged to be more difficult, we do not find the correlations with a lower standard (a judge nearest the second quintile, for example) being higher. Rather, we observe the individual standard shifting by the degree of inaccurate judgment and the correlations remaining high with the observed item difficulties near the deliberated average standard.

Moreover, in the final judgments, made from a more fully-elaborated construct, where there were differences in favor of the individual standards, the correlations of estimated difficulties and difficulties for populations near the average standard were generally high, exceeding .750. The only exceptions were in Reading, (Judge r) and (Judge g).

If Judges r and g are removed from Reading, the average Angoff standard changes less than half of a point (about .3 of a point). Therefore, even where there is a smaller convergence with the hypothesized model of the construct, it has had no functional effect on the standard.

It should be mentioned that there appears to be a judge in Writing (x), that was operating under a different construct. Again, the functional effect on the average standard of removing this judge is to raise the standard from 19.38 to 19.45, rounded to 19.5 in both cases.

Conclusion

There are many possible manipulations of the data to suggest agreement among estimated and observed item difficulties. Such agreement is a necessary component of validation of the hypothetical construct of a minimally competent group of examinees on a defined domain. In turn, validation of this type requires demonstration that the test is homogenous over the range of estimates, and, therefore, that agreement of estimates of item difficulty for the hypothetical group with observed performance of examinees near the averaged deliberated standard (across judges) is at least as high as agreement of these estimates with observed performance near each individual judge's standard.

References

AERA, APA, & NCME (1985). Standards for Educational and Psychological Testing. Washington, D.C.: author.

Angoff, W.H. (1971). Scales, Norms, and Equivalent Scores. In R.L. Thorndike (ed.), Educational Measurement. Washington, D.C., American Council on Education, pp. 514-515.

Angoff, W.H. (1984). Scales, Norms, and Equivalent Scores. Princeton, New Jersey: Educational Testing Service.

Angoff, W.H. & Modu, C.C. (1973). Equating the scales of the Prueba de Aptitud Academica and the Scholastic Aptitude Test. Research Report #3. Princeton, New Jersey: Educational Testing Service.

Brennan, R.L. & Lockwood, R.E. (1980). A comparison of the Nedelsky and Angoff cutting score procedures using generalizability theory. Applied Psychological Measurement. v. 4, no. 2, pp. 219-240.

DeMauro, G.E. (1991). Setting standards on teacher assessments by maximizing the accuracy of Angoff judgments. Paper presented at the annual conference of the National Council on Measurement in Education, Chicago.

DeMauro, G.E. & Powers, D. (1990). Internal consistency of the Angoff method of standard setting. Paper presented at the annual meeting of the National Council on Measurement in Education. Boston.

Jaeger, R.M. (1988). Establishing standards on tests used for certification of educational personnel: Validity issues. Paper presented at the annual meeting of the American Psychological Association, Atlanta.

Kane, M.T. (1986). The interpretability of passing scores. ACT Technical Bulletin Number 52. Iowa City, Iowa: The American College Testing Program.

Kerlinger, F.N. (1964). Foundations of Behavioral Research. New York: Holt, Rinehart, & Winston.

Livingston, S.A. & Zieky, M.J. (1982). Passing Scores: A Manual for Setting Standards of Performance on Educational and Occupational Tests. Princeton, New Jersey: Educational Testing Service.

References

New Jersey State Department of Education (1994). New Jersey Standard Setting Study Report of Activities. Iowa City, Iowa: National Computer Systems.

Skakun, E.N. & Kling, S. (1980). Comparability of methods for setting standards. Journal of Educational Measurement, v. 17, no. 3, pp. 229-235.

Webb, M.W. & Miller, E.R. (1995). Setting standards on constructed response items. Paper presented at the annual meeting of the National Council on Measurement in Education, San Francisco.