DOCUMENT RESUME

ED 421 533 TM 028 859

AUTHOR Bay, Luz; Nering, Michael L.
TITLE A Demonstration of Using Person-Fit Statistics in Standard Setting.
INSTITUTION ACT, Inc., Iowa City, IA.
SPONS AGENCY National Assessment Governing Board, Washington, DC.
PUB DATE 1998-04-00
NOTE 18p.; Paper presented at the Annual Meeting of the American Educational Research Association (San Diego, CA, April 13-17, 1998).
CONTRACT ZA97001001
PUB TYPE Reports - Descriptive (141) -- Speeches/Meeting Papers (150)
EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS Academic Achievement; Geography; *Goodness of Fit; High School Seniors; High Schools; *Item Response Theory; *Responses; *Standards; Test Results
IDENTIFIERS Item Score Patterns; National Assessment of Educational Progress; *Person Fit Measures; *Standard Setting

ABSTRACT

The use of person-fit methods to determine the extent to which a panelist's ratings fit the item response theory (IRT) models used in the National Assessment of Educational Progress (NAEP) is demonstrated. Person-fit methods are statistical methods that allow the identification of nonfitting response vectors. To determine whether panelists' ratings fit the IRT models used in the NAEP, the l(z) statistic (F. Drasgow, M. Levine, and E. Williams, 1985) was used. Rating data from the 1994 NAEP achievement level setting process were obtained for grade 12 geography, for which 29 panelists (primarily teachers) set levels. A response vector was created for each panelist for each achievement level using each of three p-value criteria and simulated item score string estimation (ISSE) values were created. The l(z) statistic was calculated for each of the 27 response vectors associated with each of the 29 panelists. Means and standard deviations of the l(z) distributions were computed for each cell of the experimental design and are presented in table form. Typically, they indicated that the simulated ISSE ratings or response vectors underfit the model. The results of this study provide preliminary information about the use of person-fit statistics in standard setting and are the basis for additional planned studied. (Contains 3 tables, 3 figures, and 15 references.) (SLD)

1

A Demonstration of Using Person-Fit Statistics in Standard Setting[1]

Luz Bay
Michael L. Nering

ACT

In setting achievement levels[1] for the National Assessment of Educational Progress (NAEP), there has always been an assumption that panelists' ratings fit the NAEP IRT models.[2] As stated in ACT, 1993, the NAEP Achievement Levels-Setting (ALS) process assumes that "The item response theory (IRT) analysis and scaling of the NAEP is sufficiently valid and precise so that the IRT parameter estimates can be used to map the panelists' judgments onto the NAEP scale. (p. 2)" This assumption has serious implications yet it has not been investigated by previous researchers.

What if the Panelists' Ratings Do Not Fit the Model

For the 1994 and 1996 ALS processes, achievement levels cutpoints were set by asking panelists to estimate the probability that a student performing at the borderline of each achievement level will answer each dichotomous item correctly. This item rating method is commonly referred to as the modified-Angoff method (##ref). A second method was used for polytomous items where panelists provided the average score of students performing at the borderline of each achievement level (basic, proficient, and advanced). This method is referred to as the mean estimation (ME) method (ACT, 1994).

Suppose a panelist's rating for a dichotomous item fits a linear model. This scenario is depicted in the graph in Figure 1 where the S-shaped curve is the characteristic curve for the item and the line represents the linear model that governs the panelist's rating. Suppose B is where the panelists meant to set the cutpoint, thus giving the item a r_B rating. That rating will be translated to the score scale as B', thus setting a cutpoint that is lower than where the panelist had intended.

Person-fit statistics are typically used to determine the extent to which the ability estimate for an examinee represents the underlying latent trait of interest. If an examinee has a fit index that is relatively large, then we assume that his/her response does not follow the underlying test model. Likewise, if a panelist's ratings do not follow our underlying test model, then we assume that his/her ratings do not reflect cutscores that realistic determine what is basic, proficient and advanced.

[1]Paper presented in M.L. Nering (Moderator), *Innovations in Person-fit Research*. Related papers session at the meeting of the National Council of Measurment in Education, April 14-16, 1998, San Diego, CA.

The research reported here was supported by contract #ZA97001001 from the National Assessment Governing Board to ACT.

The Use of Person Fit to Model Panelist Fit

For the 1998 NAEP in Civics and Writing, ACT proposed a new item rating method to set achievement levels cutpoints. In this method panelists are required to indicate whether a student performing at the borderline of each achievement level -- Basic, Proficient or Advanced, is likely to respond to each dichotomous item correctly (i.e., Yes="1" and No="0"). Additionally, for each polytomous item scored from 0 to *n*, panelists are required to indicate the likely score of a student performing at the borderline of each achievement level. The ratings provided by each panelist for a mix of polytomous and dichotomous items create an item score string for a student performing at the borderline of each achievement level; thus, the name item score string estimation (ISSE) method. The cutpoints can then be computed using any method that is similar to methods used for computing scale scores for students, although the interest is on group statistics (i.e., cutpoints set by a group of judges).[3]

Because each cutpoint represents the scale score of students performing at the borderline of each achievement level, a judge's ISSE ratings is a response vector of a student performing at the borderline. Since each cutpoint is computed using the same procedures that are used to compute student scores, person-fit methods may be used to determine whether a judge's ratings are congruent with the underlying IRT models. Thus, a statistic such as l_z (Drasgow, Levine, & Williams, 1985) can be determined for each judge, where:

$$l_z = \frac{l_o - E(l_o)}{\left[var(l_o)\right]^{1/2}} \qquad (1)$$

The *lo* term in Equation 1 is defined by Levine and Rubin (1979) as simply the unstandardized person-fit statistic:

$$l_o = \ln \prod_{i=1}^{n} P_i(\hat{\theta})^{u_i} Q_i(\hat{\theta})^{(1-u_i)} \qquad (2)$$

The expected value [E(*lo*)] variance [var(*lo*)] in Equation 1 can be respectively defined as:

$$E(l_o) = \sum_{i=1}^{n} \left\{ P_i(\hat{\theta}) \ln P_i(\hat{\theta}) + \left[1 - P_i(\hat{\theta}) \ln P_i(\hat{\theta})\right] \right\} \qquad (3)$$

and

$$var(l_o) = \sum_{i=1}^{n} P_i(\hat{\theta}) \left[1 - P_i(\hat{\theta})\right] \left\{ \ln \left\{ \frac{P_i(\hat{\theta})}{\left[1 - P_i(\hat{\theta})\right]} \right\} \right\}^2 \qquad (4)$$

All *Pi* values in Equations 1 through 4 represent the probability of a correct response to dichotomous item *i* given the IRT model ($Q = 1 - P$). Because there is a mix of polytomous and dichotomous items on the test the three-parameter logistic model (3PLM) as well as the generalized partial credit models (GPCM) were used. These models can be respectively defined as:

$$P(\theta) = c_i + (1-c_i)\frac{\exp\left[Da_i(\theta-b_i)\right]}{1+\exp\left[Da_i(\theta-b_i)\right]} \quad (5)$$

and

$$P_{jk}(\theta) = \frac{\exp\left[\sum_{v=1}^{k} a_j(\theta-b_j+d_v)\right]}{\sum_{c=1}^{m}\exp\left[\sum_{v=1}^{c} a_j(\theta-b_j+d_v)\right]} \quad (6)$$

Purpose of the Study

This study aims to demonstrate the use of person-fit methods to determine the extent to which a panelist's ratings fit the IRT models used in NAEP. Person-fit methods are statistical methods that will allow us to identify nonfitting-response vectors. To determine whether panelists' ratings fit the IRT models used in NAEP we will use the l_z statistic described above and using the dichotomous and polytomous panelist item strings.

P-Value Criteria

For this demonstration, rating data from 1994 NAEP ALS process for grade 12 geography were used. Because person-fit statistics use response vectors they can only be used for this type of study if the panelists' ratings are of a form similar to ISSE ratings. However, rating data from the 1994 NAEP ALS processes do not form response vectors. For this study, ISSE ratings were simulated from ME ratings by using three p-value criteria: 50%, 65%, and 80%.

The three p-value criteria have significant meanings in NAEP. For the 1994 and 1996 NAEP ALS processes, exemplar items[4] were selected so that students performing at the achievement level have, on average, at least a 50% a probability of answering the item correctly. In recent years, NAEP results have been reported using "anchor levels" and a procedure called "behavioral anchoring" has been used to determine what students can do at these levels. The descriptions of what the students can do at each anchor level were based on the items that students performing at that level have a 65% or 80% chance of responding correctly, depending on which NAEP assessment (Allen, Johnson, Mislevy, and Thomas, 1996; Johnson, Mislevy, and Thomas, 1994).

The three p-value criteria were used to simulate ISSE ratings from ME ratings provided by the 1994 NAEP ALS grade 12 geography panelists. Another objective of this study is to determine whether there is a p-value criterion that produces better fitting ISSE ratings.

Achievement Levels

ALS panelists estimate the performance of students at the borderline of each of the three achievement levels. It will be investigated whether panelists can estimate student performance at one level better than they can estimate student performance at other levels. That is, are the ratings that panelists provide for an achievement level fitting the model better than the ratings they provide for other levels? Additionally, it will be determined whether a p-value criterion produces better fitting ISSE ratings for different achievement levels.

Round of Ratings

To set achievement levels cutpoints, panelists rate the items in three rounds (ACT, 1993 – Design Document). Between rounds, panelists are provided with feedback based on the previous round of ratings and other information that they may consider in the subsequent rounds of ratings. The following is feedback provided to the panelists:

1. Cutpoints set for the grade level and the corresponding standard deviations

2. Rater Location Feedback
   A chart that provides each individual panelist where he/she set his/her cutpoints relative to the cutpoints set by other panelists. This chart indicates interrater consistency.

3. Inrarater Consistency Feedback
   Indicates how consistent a panelist's rating of each item is consistent with his/her ratings of the other items.

4. Whole Booklet Feedback
   Panelists are provided with the percent-correct score on a form of the assessment that is expected of a student performing at the borderline of each achievement level. This assessment form was the one taken by the panelists during their training.

5. P-Value Feedback
   The overall percentage of students who responded correctly to each of the dichotomous items that were rated. For each polytomous item, the p-value feedback includes the mean score on each item and the percentage of students who performed at each score level.

Panelists are told that the feedback provided is for their information and they may adjust their ratings on subsequent rounds based on this information.

Data from ALS processes conducted by ACT indicate that panelists do adjust their ratings based on the feedback and other information provided between rounds (e.g., ACT, 1997). In this study, we want to determine whether panelists provide ratings that fit the model better from round to round. That is, do panelists provide better fitting ratings after getting feedback information.

Panelist Type

It is a policy of the National Assessment Governing Board (NAGB), the agency charged by Congress to formulate policy guidelines for NAEP, that ALS panels be composed of 55% educators, 15% nonteacher educators, and 30% members of the general public. There are factions of the educational measurement community who believe that only "experts" should be called-on to set standards. Because the three types of panelists have different levels of

"expertise," some believe that some types of panelists provide more "accurate" ratings than others. It will be investigated in this study whether there are significant differences in how well the ratings provided by different types of panelists fit the IRT model. Moreover, we will investigate whether the model fit of the ratings provided by different types of panelists differ based on the p-value criterion used.

Rating Group

Each grade level panel was divided into two groups to reduce the intellectual and time demand on ALS panelists. The two item rating groups are about equivalent relative to the number of panelist belonging to each type, race or ethnicity, sex, and region of the country that the panelist represents. Each item rating group is assigned to rate about half of the assessment items. Each half is called an item rating pool. The item pool for the grade level is divided to form two item rating pools that are about equivalent with respect to item type (multiple-choice, or constructed response), subcontent area[5], and difficulty. In this study, it will be determined whether ratings provided by one group of panelists fit better than the ratings provided by the other group. A difference in the fit of the ratings provided by different group may be an indication that items in one item pool is more difficult to judge or rate than the items in the other item pool.

Method

Dataset

Rating data from the 1994 NAEP ALS process were obtained. For the purposes of this study, only data from grade 12 geography were used. There were 29 panelists who set achievement levels for grade 12 geography. There were 18 teachers, three nonteacher educators, and eight members of the general public. There were 12 male and 17 female panelists. Three of the panelists represent minorities. The 15 panelists in Group A rated 84 items (54 dichotomous and 26 constructed response items), while 14 Group B panelists rated 71 items (50 dichotomous and 21 constructed response items).

Each panelist provided three round of ratings for each achievement level, Basic, Proficient, and Advanced. During each round of ratings, each panelist estimated the percentage of students performing at the borderline of each achievement level who would answer each multiple-choice item correctly. For each constructed response item, panelists estimated the average score of students performing at the borderline of each achievement level.

During the ALS meetings, cutpoints set by each panelist for each achievement level were computed for the purpose of providing feedback. These individual cutpoints were computed in a way that was consistent with the computation of grade level cutpoints.[6] Each panelist's cutpoints on the θ scale were used for this study to serve as the ability estimate of students performing at the borderline of each achievement level; the performance of whom they are trying to estimate in the item rating process.

Determining Panelist Fit

A response vector was created for each panelist for each achievement level for each round of rating using each of the three p-value criterion (i.e., 50%, 65%, and 80%). Thus, 27 response vectors were created for each of the 29 panelists. These response vectors were created by transforming each modified-Angoff rating to a "0/1" rating. That is,

using the p-value criterion (where p = .50, .65, or .80) each modified-Angoff rating higher than p was replaced by "1" and all others by "0." The ME ratings provided by panelists for constructed response items scored 1 through 4, which were real numbers between 1 and 4, were replaced with 1, 2, 3, or 4 using the following rules: If the p-value criterion is p% and n = 1, 2, 3, or 4, then an ME rating greater than $n + p/100$ but less than n+1 is replaced with n+1. Otherwise, the rating is replaced with n. The rules for converting ME ratings are clarified in Table 1.

The l_z statistic was calculated for each of the 27 response vectors associated with each of the 29 panelists. The l_z was found by making use of Equations 1 through 6.

Analyses

The distributional characteristics of l_z statistics for various subgroups of panelists' were computed to determine how well the panelists fit the model overall. The mean and standard deviation of the l_z have expected values of 0.0 and 1.0, respectively. Thus, large differences in the mean and standard deviations would suggest overall lack of model fit. In addition to the distributional characteristics a five-factor analysis of variance (ANOVA) with repeated measures on three factors was performed using l_z as the dependent variable. The five factors (VARIABLES) were:

Panelist Type (TYPE)
Rating Group (GROUP)
Round of Ratings (ROUND)
Achievement Level (LEVEL)
P-Value Criterion (PVCRIT)

Each panelist (JUDGE) was nested within TYPE and GROUP, while the variables ROUND, LEVEL, and PVCRIT were repeated factors. A summary of the experimental design is shown in Figure 2.

The main effects of all the five factors are of interest, where the research questions to be answered by this design are:

- Is there a p-value criterion that produces ISSE type ratings that fit the NAEP IRT models better?
- Do panelists ratings for Basic, Proficient, and Advanced levels fit the NAEP IRT models equally well? Can panelists provide judgment of student performance for one level better than they can judge student performance for another level?
- Do panelists ratings improve in fit across rounds? Do panelists provide ratings that better fit the NAEP IRT models after being provided with feedback information?
- Do ratings provided by panelists from different rating groups fit the NAEP IRT models equally well? Are there any indications that items in one rating pool are relatively more difficult to rate?
- Do ratings provided by different types of panelists fit the NAEP IRT models equally well? That is, do teachers who may be considered the "expert" judges provide ratings that better fit the models?

Moreover, it will be determined whether there were interaction effects between TYPE and PVCRIT, and LEVEL and PVCRIT. The research questions that correspond to these interaction effects are:

- Is the p-value criterion that produces better fitting ISSE ratings different for different types of panelists?
- Is the p-value criterion that produces better fitting ISSE ratings different for different achievement

levels?

The ANOVA was performed using the SAS software. Because the cell sizes were unequal, generalized linear models (GLM) procedures were used.

Results

The means and standard deviations of the l_z distributions were computed for each cell of the experimental design. These means and standard deviations are presented in Table 2. Notice that typically the l_z values indicated that the simulated ISSE ratings or response vectors underfit the model. That is, the negative values in Table 2 suggest that the panelists were responding in a manner that was inconsistent with the test model. For example, in Round 1, Basic level using the .50 criterion for Teachers the mean l_z value was -3.61 (SD=5.43). There did not appear to be an improvement in model fit for this group of panelists (at the same proficiency level and p-value criterion) in subsequent rounds (Round 2 mean = -4.62; Round 3 mean = -3.73). Interestingly, the largest negative l_z values were typically found under the .80 p-value criterion (particularly in the proficient and advance achievement levels). The most promising values in Table 2 (i.e., where the values were closest to their expected values) were found in the advanced levels using the .50 p-value criterion. For example, in Round 1 for panelists in Group B the mean and SD were -0.09 and 0.94, respectively.

A summary of the ANOVA is presented in Table 3. To answer the first research question, consider the test on the effect of PVCRIT on l_z. Clearly, with F=194.41 (p=0.0001), there is a significant difference between the average l_z values obtained using different p-value criterion. The overall (unadjusted) means indicate that using a p-value criterion of 50% provide better fitting ISSE ratings. The averages of the l_z values are -1.88, -3.92, and -9.16 for p-value criteria of 50%, 65%, and 80%, respectively. This main effect is also clearly noticeable by visual inspection of Table 2.

Significant differences between l_z values for different achievement levels were also found (F=14.79, p=0.0001). It seems that panelists' ratings for the Advanced level produced l_z values that were closer to 0 than the other levels . The overall (unadjusted) means of l_z for the Basic, Proficient, and Advanced levels were -4.39, -7.78, and -2.81, respectively.

The main effect of ROUND (F=13.66, p=0.0001) indicate that there was improvement in model fit of the ratings that the panelists provide across rounds of ratings. First round ratings have (unadjusted) average l_z of -5.93. For rounds 2 and 3, the respective averages of the l_zs were -4.85 and -4.19. This may be considered an indication that panelists take into consideration the feedback information provided to them between rounds of ratings, and that such information make them better judges of student performance.

There were no main effects from factors rating group and panelist type, with F=0.43 (p=0.5286) and F=0.99 (p=0.3851), respectively. The absence of GROUP effect is indicate the lack of reason to believe that the item rating pools are not equivalent with respect to the rating challenge posed by the items. The absence of TYPE effect on l_z seem to indicate that with respect to providing item ratings, classroom teachers are not any better than nonteacher educators or members of the general public. As a matter of fact, nonteacher educators provided slightly better fitting ratings in this case.

A hypothesis is that different p-value criteria will produce better fitting ratings for different types of panelists. For example, a higher p-value criterion might produce better ratings for teachers considering that they are the "expert" judges. Studies show, that standard setting judges with higher levels of expertise tend to provide higher ratings due to their higher expectations (##ref). The absence interaction effect between PVCRIT and TYPE seem to indicate that the hypothesis is false.

Another interaction of interest is that of PVCRIT and LEVEL. Clearly, there is interaction between these two factors (F=34.12, p=0.0001). A graphical representation of the interaction of these factors is in Figure 3. Unadjusted means of the l_z were used to illustrate the interaction. Changing the p-value criterion does not seem to affect the fit of the Basic ratings. Using 65% instead of 50% drastically affects the fit of Proficient ratings, but it does not seem to affect the fit of Advanced ratings as much. Using 80% instead of 50% or 65% drastically affects the fit of both Proficient and Advanced ratings.

Discussion

This study is a first step in making use of a person-fit statistic in standard setting. The results of this study were promising in that it provides information that are valuable to standard setters and policy makers. Analyses performed in this study will be repeated on the 1994 NAEP Geography ALS rating data for grades 4 and 8 and also for 1994 NAEP U.S. History and 1996 NAEP Science for all grade levels. Results will be compared across all subject areas and grade levels.

If ISSE ratings will be used for the setting achievement levels for the 1998 NAEP for Civics and Writing[7], person-fit statistics will be used to determine whether panelists' ratings fit the IRT models used. This will provide indications whether the assumption that panelists' ratings are governed by the same model that governs student responses is reasonable.

References

ACT (1997). *Setting achievement levels on the 1996 National Assessment of Educational Progress in Science final report.* Iowa City, IA: Author.

ACT (1995a). *NAEP reading revisit: An evaluation of the 1992 achievement levels descriptions.* Iowa City, IA: Author.

ACT (1995b). *Results of the 1994 Geography NAEP Achievement Levels-Setting pilot study.* Iowa City, IA: Author.

ACT (1995c). *Results of the 1994 U.S. History NAEP Achievement Levels-Setting pilot study.* Iowa City, IA: Author.

ACT (1994). *Design document for setting achievement levels on the 1994 National Assessment of Educational Progress in Geography and U.S. History and the 1996 National Assessment of Educational Progress in Science.* Iowa City, IA: Author.

ACT (1993a). *Setting achievement levels on the 1992 National Assessment of Educational Progress in Mathematics, Reading, and Writing: A technical report on reliability and validity.* Iowa City, IA: Author.

Allen, N.L., Johnson, E.G., Mislevy, R.J., & Thomas, N. (1996). *Scaling procedures.* In N.L. Allen, D.L. Kline, & C.A. Zelenak, *The NAEP 1994 Technical Report*, Wahington D.C: National Center for Education Statistics.

Johnson, E.G., Mislevy, R.J., & Thomas, N. (1994). *Scaling procedures.* In E.G. Johnson & J.E. Carlson, *The NAEP 1992 Technical Report*, Wahington D.C: National Center for Education Statistics.

Bay, L. (1997). Comparing student performance on different item formats relative to achievement levels cutpoints. In M.L. Bourque (Moderator), *Setting Standards for NAEP.* Related-papers session conducted at the meeting of the National Council on Educational Measurement, April, 1998, San Diego, CA.

Bay, L., Chen, L., & Reckase, M. D. (1997). *The grid: A possible rating method for the 1998 NAEP writing achievement levels-setting process.* A report prepared for the meeting of the 1998 NAEP Achievement Levels-Setting Project Technical Advisory Committee for Standard Setting (TACSS), Oct. 2-3, 1997, St. Louis, MO.

Bay, L., & Hanson, B. A. (1997). *Computing achievement levels cutpoints from NAEP booklet classification studies: A secondary analysis.* A report prepared for the meeting of the 1998 NAEP Achievement Levels-Setting Project Technical Advisory Committee for Standard Setting (TACSS), Oct. 2-3, 1997, St. Louis, MO.

Chen, Wen-Hung (1997). *Setting achievement levels for NAEP using item score string: A simulation study.* A report prepared for the meeting of the 1998 NAEP Achievement Levels-Setting Project Technical Advisory Committee for Standard Setting (TACSS), Oct. 2-3, 1997, St. Louis, MO.

Drasgow, F., Levine, M. V., & Williams, E. A. (1985). Appropriateness measurement with polychotomous item response models and standardized indices. *British Journal of Mathematical and Statistical Psychology, 38,* 67-86.

Levine, M. V., & Rubin, D. B. (1979). Measuring the appropriateness of multiple-choice test scores. *Journal of Educational Statistics, 4,* 269-290.

Table 1
Conversion of Polytomous Item Ratings

Converted Ratings	P-Value Criterion		
	50	65	80
1	1.00-1.50	1.00-1.65	1.00-1.80
2	1.51-2.50	1.66-2.65	1.81-2.80
3	2.51-3.50	2.66-3.65	2.81-3.80
4	3.51-4.00	3.66-4.00	3.81-4.00

Table 2

Means and Standard Deviations of the l_z Statistics, by Panelist Type and by Grade

Round	Level	P-Value Criterion	Panelist Type						Rating Group			
			Teacher (n=18)		Nonteacher Educator (n=3)		General Public (n=8)		A (n=15)		B (n=14)	
			Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD
1	Basic	50	-3.61	5.43	-4.86	7.51	-3.63	4.86	-2.16	5.13	-5.45	5.09
		65	-3.50	5.58	-5.31	7.86	-4.17	5.62	-1.86	5.00	-6.02	5.60
		80	-3.32	6.14	-5.03	8.00	-4.77	6.35	-1.29	5.03	-6.70	6.21
	Proficient	50	-4.12	2.70	-2.37	2.77	-3.38	3.75	-5.29	2.97	-2.07	1.92
		65	-9.14	4.14	-4.27	4.62	-9.35	3.84	-11.17	2.35	-6.05	4.25
		80	-15.01	4.40	-9.40	0.94	-14.7	5.15	-13.17	3.68	-15.63	5.30
	Advanced	50	-0.28	0.75	0.16	0.88	0.08	0.77	-0.17	0.58	-0.09	0.94
		65	-2.35	3.73	-0.60	1.97	-1.97	2.87	-3.78	3.86	-0.22	0.92
		80	-13.19	7.11	-7.52	0.34	-14.0	9.00	-19.02	3.53	-6.23	5.42
2	Basic	50	-4.62	4.73	-3.77	1.75	-5.00	2.72	-5.12	4.10	-4.12	3.92
		65	-4.84	4.71	-5.31	3.64	-5.79	3.10	-5.09	4.01	-5.23	4.38
		80	-5.04	4.73	-5.19	4.08	-6.41	2.59	-4.76	3.91	-6.16	4.32
	Proficient	50	-1.58	1.82	-0.34	0.49	-1.20	1.69	-1.46	2.18	-1.22	1.02
		65	-6.55	4.59	-1.16	0.63	-7.26	5.46	-8.65	4.55	-3.55	3.68
		80	-16.18	4.21	-12.2	5.05	-14.2	4.25	-15.66	3.13	-14.79	5.45
	Advanced	50	0.30	0.75	0.58	0.17	0.76	0.62	0.87	0.53	0.01	0.57
		65	-0.09	0.81	0.40	0.02	0.04	1.19	0.23	0.95	-0.25	0.75
		80	-6.96	5.26	-2.52	1.49	-5.57	4.59	-8.89	5.20	-3.14	2.11
3	Basic	50	-3.73	3.93	-1.53	0.77	-4.23	2.68	-4.38	3.88	-2.85	2.81
		65	-4.25	4.01	-3.64	2.65	-5.02	2.93	-4.34	3.73	-4.47	3.49
		80	-4.48	4.04	-4.06	4.11	-5.60	2.65	-3.98	3.59	-5.57	3.61
	Proficient	50	-1.12	1.37	0.03	0.23	-0.54	1.68	-0.95	1.77	-0.73	0.96
		65	-5.63	4.31	-0.61	0.63	-5.72	4.81	-7.39	4.31	-2.73	3.15
		80	-15.19	3.70	-11.5	4.77	-13.9	4.45	-14.92	2.65	-13.98	5.18
	Advanced	50	0.46	0.65	0.64	0.31	0.96	0.60	1.07	0.40	0.13	0.45
		65	0.09	0.66	0.42	0.39	0.35	0.91	0.46	0.76	-0.08	0.54
		80	-6.25	4.96	-2.35	1.07	-4.51	3.69	-7.99	4.86	-2.56	1.26

Table 3

Analysis of Variance Summary Table

Source	df	SS	MS	F	p
Among Judges					
TYPE	2	201.04	100.52	0.99	0.3851
GROUP	1	43.43	43.43	0.43	0.5286
TYPE*GROUP	2	43.37	21.69	0.21	0.8084
JUDGE	23		101.04		
Within Judges					
ROUND	2	235.52	117.76	13.66	0.0001
ROUND*TYPE	4	3.09	0.77	0.09	0.9852
ROUND*GROUP	2	0.05	0.02	0.00	0.9971
ROUND*TYPE*GROUP	4	2.17	0.54	0.06	0.9925
ROUND*JUDGE	46	396.61	8.62		
LEVEL	2	1604.59	802.30	14.79	0.0001
LEVEL*TYPE	4	150.31	37.58	0.69	0.6008
LEVEL*GROUP	2	759.32	379.66	7.00	0.0022
LEVEL*TYPE*GROUP	4	171.52	42.88	0.79	0.5374
LEVEL*JUDGE	46	2495.37	54.24		
PVCRIT	2	3761.76	1880.88	194.41	0.0001
PVCRIT*TYPE	4	36.16	9.04	0.93	0.4525
PVCRIT*GROUP	2	39.87	19.94	2.06	0.1390
PVCRIT*TYPE*GROUP	4	55.10	13.77	1.42	0.2412
PVCRIT*JUDGE	46	445.05	9.67		
ROUND*LEVEL	4	249.94	62.49	7.11	0.0001
ROUND*LEVEL*TYPE	8	6.77	0.85	0.10	0.9993
ROUND*LEVEL*GROUP	4	4353.61	108.90	12.39	0.0001
ROUND*LEVEL*TYPE*GROUP	8	62.24	7.78	0.89	0.5323
ROUND*LEVEL*JUDGE	92	808.76	8.79		
ROUND*PVCRIT	4	9.21	2.30	1.64	0.1708
ROUND*PVCRIT*TYPE	8	19.32	2.41	1.72	0.1039
ROUND*PVCRIT*GROUP	4	2.51	0.63	0.45	0.7749
ROUND*PVCRIT*TYPE*GROUP	8	13.23	1.65	1.18	0.3206
ROUND*PVCRIT*JUDGE	92	129.13	1.40		
LEVEL*PVCRIT	4	1764.18	441.04	34.12	0.0001
LEVEL*PVCRIT*TYPE	8	84.17	10.52	0.81	0.5921
LEVEL*PVCRIT*GROUP	4	521.79	130.45	10.09	0.0001
LEVEL*PVCRIT*TYPE*GROUP	8	69.69	8.71	0.67	0.7132
LEVEL*PVCRIT*JUDGE	92	1189.14	12.93		
ROUND*LEVEL*PVCRIT	8	390.26	48.78	18.78	0.0001
ROUND*LEVEL*PVCRIT*TYPE	16	26.35	1.65	0.63	0.8534
ROUND*LEVEL*PVCRIT*GROUP	8	218.46	27.31	10.52	0.0001
ROUND*LEVEL*PVCRIT*TYPE*GROUP	16	71.48	4.48	1.72	0.0461
ROUND*LEVEL*PVCRIT*JUDGE	184	477.83	2.60		

Figure 1
A Possible Scenario of a Panelist's Rating not Fitting the IRT Model

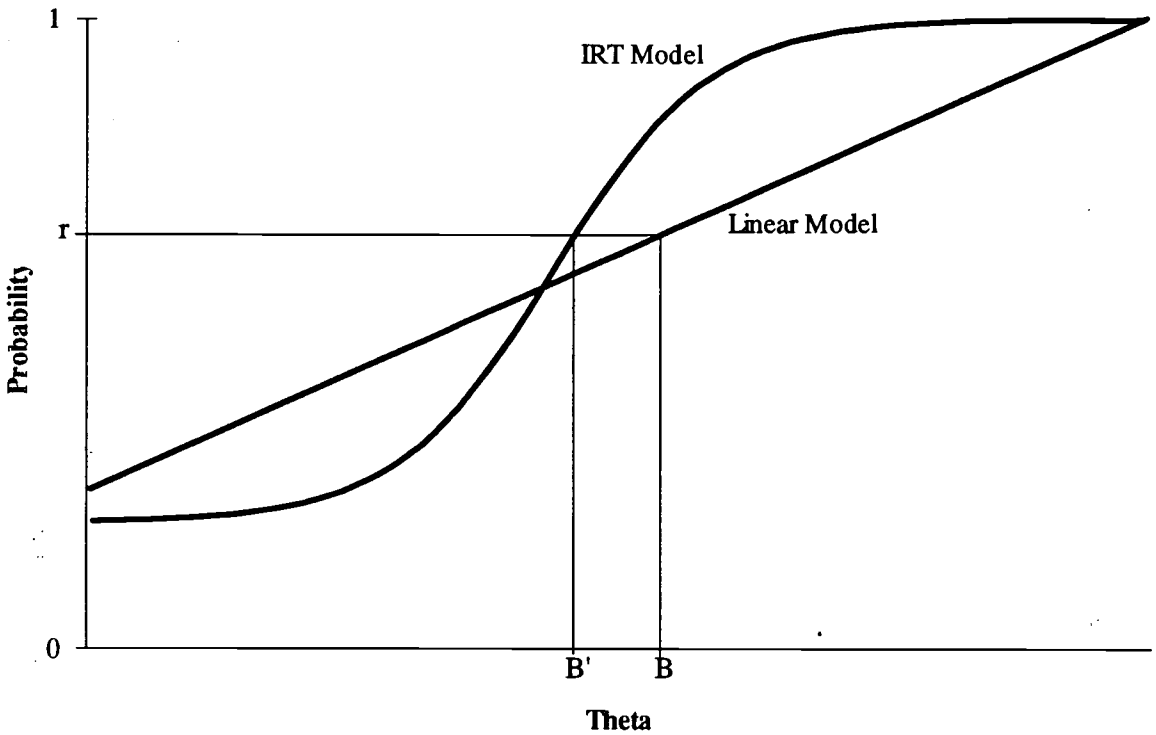


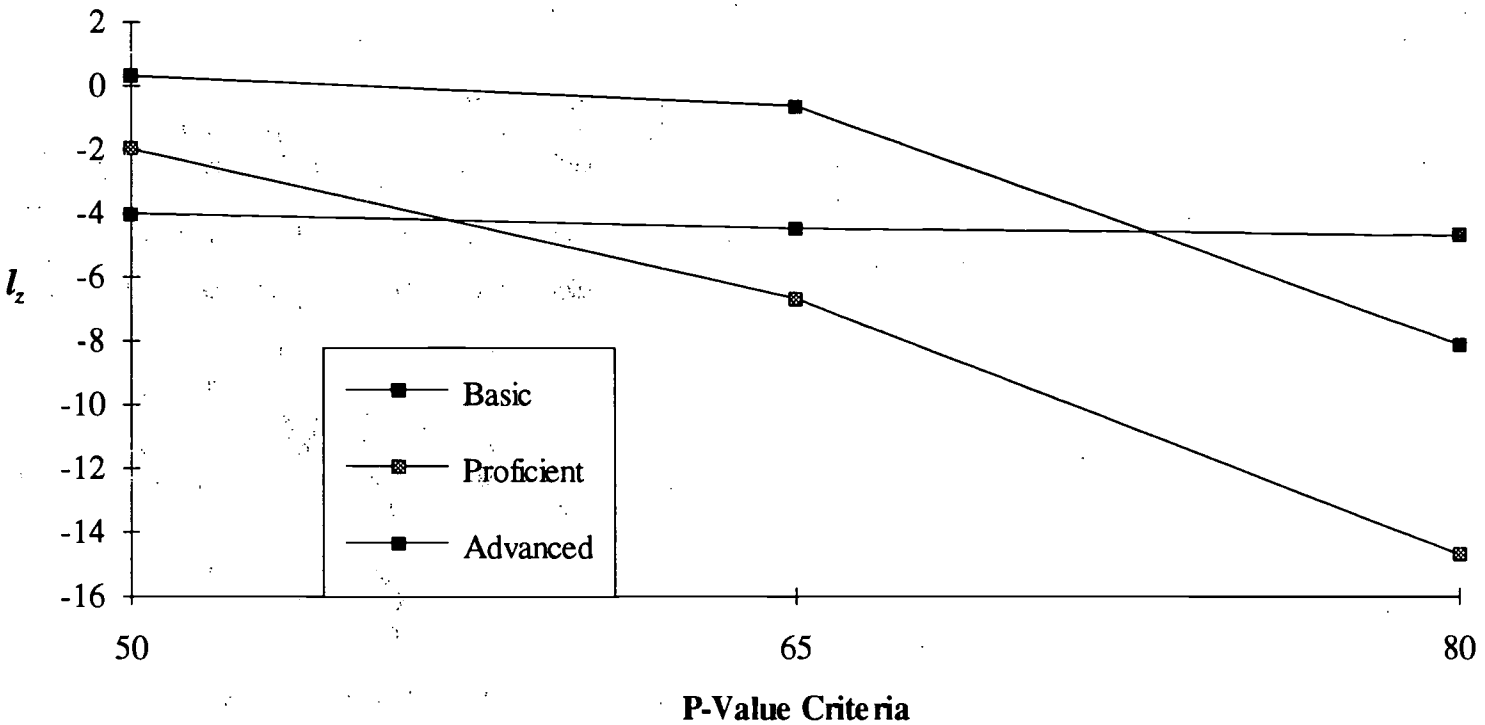
Figure 2

Experimental Design

		ROUND																										
		1									2									3								
LEVEL:		Basic			Proficient			Advanced			Basic			Proficient			Advanced			Basic			Proficient			Advanced		
PVCRIT:		50	65	80	50	65	80	50	65	80	50	65	80	50	65	80	50	65	80	50	65	80	50	65	80	50	65	80
GROUP	TYPE																											
A	1	G_1	G_1	G_1	G_1	G_1	G_1	G_1	G_1	G_1	G_1	G_1	G_1	G_1	G_1	G_1	G_1	G_1	G_1	G_1	G_1	G_1	G_1	G_1	G_1	G_1	G_1	G_1
A	2	G_2	G_2	G_2	G_2	G_2	G_2	G_2	G_2	G_2	G_2	G_2	G_2	G_2	G_2	G_2	G_2	G_2	G_2	G_2	G_2	G_2	G_2	G_2	G_2	G_2	G_2	G_2
A	3	G_3	G_3	G_3	G_3	G_3	G_3	G_3	G_3	G_3	G_3	G_3	G_3	G_3	G_3	G_3	G_3	G_3	G_3	G_3	G_3	G_3	G_3	G_3	G_3	G_3	G_3	G_3
B	1	G_4	G_4	G_4	G_4	G_4	G_4	G_4	G_4	G_4	G_4	G_4	G_4	G_4	G_4	G_4	G_4	G_4	G_4	G_4	G_4	G_4	G_4	G_4	G_4	G_4	G_4	G_4
B	2	G_5	G_5	G_5	G_5	G_5	G_5	G_5	G_5	G_5	G_5	G_5	G_5	G_5	G_5	G_5	G_5	G_5	G_5	G_5	G_5	G_5	G_5	G_5	G_5	G_5	G_5	G_5
B	3	G_6	G_6	G_6	G_6	G_6	G_6	G_6	G_6	G_6	G_6	G_6	G_6	G_6	G_6	G_6	G_6	G_6	G_6	G_6	G_6	G_6	G_6	G_6	G_6	G_6	G_6	G_6

Figure 3

Interaction of Achievement Level and P-Value Criterion



Endnotes

1. Achievement Levels

As authorized by law, the National Assessment Governing Board has adopted student achievement levels for reporting NAEP results. These levels represent an informed judgment of "how good is good enough" on NAEP -- key information not available for simple averages. The three levels are used as the primary means of reporting what students should know and be able to do on the National Assessment.

Basic This level denotes partial mastery of prerequisite knowledge and skills that are fundamental for proficient work at each grade.

Proficient This level represents solid academic performance for each grade assessed. Students reaching this level have demonstrated competency over challenging subject matter, including subject-matter knowledge, application of such knowledge to real world situations, and analytical skills appropriate to the subject matter.

Advanced This level signifies superior performance.

2. For descriptions of the NAEP Achievement Levels-Setting processes please refer to ACT, 1994, and ACT, 1997.

3. Simulation studies to determine the best computational method are reported in Chen, 1997.

4. The ALS process has three products: (1) the achievement levels descriptions state what students should know and be able to do at each level; (2) the achievement levels cutpoints are the score on the NAEP performance scale that separate adjacent achievement levels; and, (3) exemplar items are items in the assessment that are illustrative of what students can do at each achievement level.

5. Each item in the NAEP geography assessment is classified into one of the three subcontent areas of geography that are specified in the framework (NAGB, 1993). These areas are: (1) Space and Place; (2) Environment and Society; and, (3) Spatial Dynamics and Connections.

6. For a concise description of how grade level cutpoints were computed for the 1994 and 1996 NAEP ALS processes, see Bay, 1997.

7. Two other rating methods are being considered for the NAEP writing ALS process. These methods are described in Bay and Hanson, 1997 and Bay, Chen, and Reckase, 1997.

U.S. Department of Education
Office of Educational Research and Improvement (OERI)
National Library of Education (NLE)
Educational Resources Information Center (ERIC)

ERIC

TM028859

REPRODUCTION RELEASE

(Specific Document)

I. DOCUMENT IDENTIFICATION:

Title: A Demonstration of Using Person-Fit Statistics in Standard Setting	
Author(s): Luz Bay, Michael L. Nering	
Corporate Source: ACT, Inc.	Publication Date:

II. REPRODUCTION RELEASE:

In order to disseminate as widely as possible timely and significant materials of interest to the educational community, documents announced in the monthly abstract journal of the ERIC system, *Resources in Education* (RIE), are usually made available to users in microfiche, reproduced paper copy, and electronic media, and sold through the ERIC Document Reproduction Service (EDRS). Credit is given to the source of each document, and, if reproduction release is granted, one of the following notices is affixed to the document.

If permission is granted to reproduce and disseminate the identified document, please CHECK ONE of the following three options and sign at the bottom of the page.

The sample sticker shown below will be affixed to all Level 1 documents	The sample sticker shown below will be affixed to all Level 2A documents	The sample sticker shown below will be affixed to all Level 2B documents
PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL HAS BEEN GRANTED BY Sample TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC) 1	PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL IN MICROFICHE, AND IN ELECTRONIC MEDIA FOR ERIC COLLECTION SUBSCRIBERS ONLY, HAS BEEN GRANTED BY Sample TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC) 2A	PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL IN MICROFICHE ONLY HAS BEEN GRANTED BY Sample TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC) 2B
Level 1	Level 2A	Level 2B
X	☐	☐
Check here for Level 1 release, permitting reproduction and dissemination in microfiche or other ERIC archival media (e.g., electronic) *and* paper copy.	Check here for Level 2A release, permitting reproduction and dissemination in microfiche and in electronic media for ERIC archival collection subscribers only	Check here for Level 2B release, permitting reproduction and dissemination in microfiche only

Documents will be processed as indicated provided reproduction quality permits.
If permission to reproduce is granted, but no box is checked, documents will be processed at Level 1.

I hereby grant to the Educational Resources Information Center (ERIC) nonexclusive permission to reproduce and disseminate this document as indicated above. Reproduction from the ERIC microfiche or electronic media by persons other than ERIC employees and its system contractors requires permission from the copyright holder. Exception is made for non-profit reproduction by libraries and other service agencies to satisfy information needs of educators in response to discrete inquiries.

Sign here, → please

Signature:	Printed Name/Position/Title: Luz Bay/Psychometrician
Organization/Address: ACT, 2201 N. Dodge St., PO Box 168 Iowa City, IA 52243-0168	Telephone: 319-337-1639 FAX: 319-337-1497
	E-Mail Address: Bay@act.org Date: 5-11-98

(over)

III. DOCUMENT AVAILABILITY INFORMATION (FROM NON-ERIC SOURCE):

If permission to reproduce is not granted to ERIC, or, if you wish ERIC to cite the availability of the document from another source, please provide the following information regarding the availability of the document. (ERIC will not announce a document unless it is publicly available, and a dependable source can be specified. Contributors should also be aware that ERIC selection criteria are significantly more stringent for documents that cannot be made available through EDRS.)

Publisher/Distributor:
Address:
Price:

IV. REFERRAL OF ERIC TO COPYRIGHT/REPRODUCTION RIGHTS HOLDER:

If the right to grant this reproduction release is held by someone other than the addressee, please provide the appropriate name and address:

Name:
Address:

V. WHERE TO SEND THIS FORM:

Send this form to the following ERIC Clearinghouse:

THE UNIVERSITY OF MARYLAND
ERIC CLEARINGHOUSE ON ASSESSMENT AND EVALUATION
1129 SHRIVER LAB, CAMPUS DRIVE
COLLEGE PARK, MD 20742-5701
Attn: Acquisitions

However, if solicited by the ERIC Facility, or if making an unsolicited contribution to ERIC, return this form (and the document being contributed) to:

ERIC Processing and Reference Facility
1100 West Street, 2nd Floor
Laurel, Maryland 20707-3598

Telephone: 301-497-4080
Toll Free: 800-799-3742
FAX: 301-953-0263
e-mail: ericfac@inet.ed.gov
WWW: http://ericfac.piccard.csc.com