NATIONAL CENTER FOR EDUCATION STATISTICS

Working Paper Series

U. S. Department of Education
Office of Educational Research and Improvement

NATIONAL CENTER FOR EDUCATION STATISTICS

Working Paper Series

Optimal Rating Procedures and Methodology for NAEP Open-ended Items

Working Paper No. 97-37 November 1997

Contact: Steven Gorman
Assessment Group
(202) 219-1937
e-mail: steven_gorman@ed.gov
or sgorman@inet.ed.gov

U. S. Department of Education
Office of Educational Research and Improvement

U.S. Department of Education
Richard W. Riley
Secretary

Office of Educational Research and Improvement
Ricky T. Takai
Acting Assistant Secretary

National Center for Education Statistics
Pascal D. Forgione, Jr.
Commissioner

Assessment Group
Gary W. Phillips
Associate Commissioner

The National Center for Education Statistics (NCES) is the primary federal entity for collecting, analyzing, and reporting data related to education in the United States and other nations. It fulfills a congressional mandate to collect, collate, analyze, and report full and complete statistics on the condition of education in the United States; conduct and publish reports and specialized analyses of the meaning and significance of such statistics; assist state and local education agencies in improving their statistical systems; and review and report on education activities in foreign countries.

NCES activities are designed to address high priority education data needs; provide consistent, reliable, complete, and accurate indicators of education status and trends; and report timely, useful, and high quality data to the U.S. Department of Education, the Congress, the states, other education policymakers, practitioners, data users, and the general public.

We strive to make our products available in a variety of formats and in language that is appropriate to a variety of audiences. You, as our customer, are the best judge of our success in communicating information effectively. If you have any comments or suggestions about this or any other NCES product or report, we would like to hear from you. Please direct your comments to:

National Center for Education Statistics
Office of Educational Research and Improvement
U.S. Department of Education
555 New Jersey Avenue, NW
Washington, DC 20208

Suggested Citation

U.S. Department of Education. National Center for Education Statistics. *Optimal Rating Procedures and Methodology for NAEP Open-ended Items*, Working Paper No. 97-37, by Richard J. Patz, Mark Wilson, and Machteld Hoskens. Project Officer, Steven Gorman. Washington, D.C.: 1997.

November 1997

Foreword

Each year a large number of written documents are generated by NCES staff and individuals commissioned by NCES which provide preliminary analyses of survey results and address technical, methodological, and evaluation issues. Even though they are not formally published, these documents reflect a tremendous amount of unique expertise, knowledge, and experience.

The ***Working Paper Series*** was created in order to preserve the information contained in these documents and to promote the sharing of valuable work experience and knowledge. However, these documents were prepared under different formats and did not undergo vigorous NCES publication review and editing prior to their inclusion in the series. Consequently, we encourage users of the series to consult the individual authors for citations.

To receive information about submitting manuscripts or obtaining copies of the series, please contact Ruth R. Harris at (202) 219-1831 or U.S. Department of Education, Office of Educational Research and Improvement, National Center for Education Statistics, 555 New Jersey Ave., N.W., Room 400, Washington, D.C. 20208-5654.

Samuel S. Peng
Acting Director
Statistical Standards and Services Group

This page intentionally left blank.

Optimal Rating Procedures and Methodology for NAEP Open-ended Items

Prepared by:

Richard J. Patz
CTB/McGraw-Hill

Mark Wilson
Machteld Hoskens
University of California at Berkelely

Prepared for:

U.S. Department of Education
Office of Educational Research and Development
National Center for Education Statistics

November 1997

The work reported herein is supported under the National Assessment of Educational Progress (CFDA No. 84.902A) as administered by the Office of Educational Research and Improvement, U.S. Department of Education. The authors thank the staff of the National Center for Education Statistics for providing the relevant data and supporting documentation.

Table of Contents

Optimal rating procedures and methodology for NAEP open-ended items

Richard J. Patz
CTB/McGraw-Hill
Monterey, CA

Mark Wilson
University of California
Berkeley, CA

Machteld Hoskens
University of California
Berkeley, CA

October 24, 1997

1 Executive summary

The National Assessment of Educational Progress (NAEP) collects data—repeated, discrete measures (test items) with hierarchical structure for both the measures and subjects (students)—that is complex by any standard. This complexity has been managed through a "divide and conquer" approach of isolating and evaluating sources of variability one at a time, using a sequence of relatively simple analyses (Patz, 1996). The cost of this simplicity for NAEP has been limits on the propagation of information from one sub-analysis to another. This has made some questions that are relatively straightforward to address in standard circumstances, quite difficult to address in NAEP. In the present study we consider NAEP's fragmented analysis of errors in the rating of open-ended responses, we develop methodology for more unified analyses, we apply the methodology to analyze rater effects in NAEP data, we investigate how to minimize rater effects using modern imaging technology, and we draw conclusions and make recommendations in light of these analyses and other analyses available in the literature.

1.1 Rater effects and what we can do about them

Raters make mistakes, and the systematic consequences of these mistakes—called rater effects—can have serious consequences for the reported results of educational tests and assessments. To complement our analyses of rater effects in NAEP, we review several recent analyses of rater effects in other programs.

A review of the literature reveals that rater effects can be quite significant, and that they may take several forms. We say rater bias is present when individual raters have consistent tendencies to be differentially severe or lenient in rating particular test items. Raters may also drift, becoming more severe or lenient over the course of a rating period. The magnitude of rater effects and their impact on test scores can be quite significant, and yet this may be well hidden when only a few traditional measures of reliability (e.g., percent exact agreement among raters) are reported. That is, it is quite possible to have high percentages of exact agreement between raters and yet have significant amounts of rater bias affecting test scores.

Providing raters with periodic feedback during the rating process can significantly improve the quality of ratings, although effective intervention requires fast and accurate algorithms for quantifying rater severity.

1.2 Analyses of rater effects in NAEP data

Analyses of data from 1992 and 1994 NAEP State Reading Assessments at grade 4 reveal several important facts about rater effects in NAEP. Rater effects, in particular, differential severity of raters scoring individual items, are detectable in NAEP. Quantifying the size and impact of these effects is hampered by several factors, two of the most important being that 1) the technology for generalizing NAEP's scaling models to include rater parameters is currently in its formative stages, and 2) the NAEP design for the allocation of responses to raters is unbalanced. Our analyses address and partially overcome the first limitation; the second limitation can and should be addressed in the design of future NAEP scoring sessions.

The within-year rater effects we detect in NAEP are not particularly large, especially when considered in light of other sources of uncertainty and error in NAEP. In the context of NAEP, these rater effects are mitigated by 1) the presence of multiple-choice items in addition to constructed-response items, 2) the randomization of individual responses to raters, and 3) the aggregate nature of NAEP's reported statistics. In this context, the across-year rater effects may be of more importance.

1.3 Optimal allocation procedures

The method of distributing responses to raters can have very significant consequences for the impact of rater errors. We found that randomization of raters to individual responses instead of intact booklets may lead to a significant reduction in the error associated with estimated proficiencies. This improvement is especially significant in the presence of large rater biases that tend to be consistent across the items of a test. This item-by-item randomization, not used in 1992 NAEP but adopted for 1994 NAEP, leads to an improvement in the accuracy of plausible values that we estimate to be equivalent to adding one additional test item to NAEP's roughly 20-item test booklets.

We propose and investigate a *stratified* randomization procedure that attempts to cancel the residual rater biases at a test score (or plausible values) level. This procedure, which could be incorporated into an integrated system for rater training, monitoring, and feedback, is shown in simulations to significantly improve proficiency estimation in the presence of severe rater effects. This finding is of general interest to the educational measurement field and should be investigated further and tested on a pilot basis.

The randomization needs to be carried out in a way that ensures that unbalanced designs do not result. Regardless of which particular randomization procedure is used, the distribution of responses to raters should be conducted in a statistically balanced fashion.

1.4 Using information from second ratings

NAEP rescores 25% of the responses to open-ended items. Currently, information from the second ratings is used only for quality control purposes. Once levels of exact agreement between ratings are deemed acceptably high, the second rating is discarded and the first is retained and used for subsequent inference (see, e.g., Johnson, Mazzeo, and Kline, 1994, pp. 88–91). Information from the second set of ratings, if incorporated appropriately, should bring greater precision to NAEP's reported statistics. In generalizability theory, the inclusion of second ratings is a standard and accepted practice. The current methods for using second ratings in item response theory (IRT) have been criticized on the grounds that they overestimate the contribution of the repeated measures (Patz, 1996). The amount of additional information available to NAEP but not used should motivate useful development of appropriate statistical methodology for incorporating information from multiple ratings of student work.

1.5 Recommendations

Based on the analyses conducted in this project, a review of related literature, and experiences from related research projects on rater effects, we make the following recommendations for consideration by the National Assessment Governing Board in its redesign of NAEP.

1. The National Center for Education Statistics (NCES) and NAEP should continue to develop a better framework for reporting on rater reliability in IRT contexts. In particular, NCES should require that NAEP contractors quantify how reported statistics would be expected to vary over replications of the professional scoring process.

2. NCES and its NAEP contractors should make more detailed information on the scoring process available, including time-stamped scoring data, read-behind, and/or check-sets data. This will facilitate investigation of the behavior of raters over the course of the scoring sessions and also from year to year.

3. NCES and its NAEP contractors should continue to develop and deploy systems that take full advantage of imaging technology in professional scoring. In particular, continued advances should be encouraged in systems for randomizing responses to raters with balanced designs, systems for monitoring rater performance, and systems for providing raters real-time feedback.

4. NCES should experiment with advanced randomization procedures based on real-time monitoring of rater severities in order to cancel residual differences in rater severities at the scale score (i.e., plausible values) level.

5. NCES should investigate improved methods of rubric standardization using imaging in order to increase the validity of NAEP's longitudinal equating.

6. NCES should encourage research to develop appropriate statistical methodology for incorporating information from multiple ratings of student work when item response theory scoring is used.

The remainder of this report provides more detail on the topics summarized above.

2 Introduction

Item response theory (IRT), introduced into NAEP analyses in the first redesign (Jones, 1996), gave NAEP much greater flexibility and more precise measurement. NAEP analyses now incorporate variability due to uncertain item characteristics (through IRT estimation of item parameters), due to sampling of students (through jackknife estimation of a sampling variance component), and due to measurement of individual proficiencies (through multiple-imputation or "plausible values" methodology).

These careful, IRT-based analyses of NAEP are presently informing steps toward simplification of early NAEP reports (Forsyth, Hambleton, Linn, Mislevy, and Yen, 1996). Because NAEP has performed careful analyses using a nearly exhaustive conditioning model, NAEP researchers may now make intelligent decisions about how to use smaller conditioning models and simpler methods for providing early NAEP results. Similarly, reporting NAEP scores in an observed score ("market-basket") metric will facilitate quicker analyses using some tools of classical test theory and generalizability theory. Valid inferences based on such simplifications are possible only because IRT plays a pivotal role in the construction of parallel market-baskets and because IRT allows us to report scores on one market-basket when items from another were administered.

In the present study we bring the rating of open-ended items directly into NAEP's existing IRT methodology. Our analyses are intended to both recognize an inherent complexity and provide a research basis for valid simplification. An IRT analysis of NAEP rater effects helps explain how the characteristics of students, items, and raters interact in the formation of NAEP open-ended item responses, and this information sheds light on the relative efficacy of simpler real-time algorithms for monitoring and controlling rater effects.

NAEP's current analysis of the rating process for open-ended items stands in contrast to its careful analyses of other sources of variance. Errors introduced into NAEP inferences due to rating errors are largely ignored in NAEP analyses (Patz, 1996). Existing analyses of NAEP rater agreement (e.g., Johnson, Mazzeo, and Kline, 1994) are limited in scope to percent agreement and limited in practice to controlling rater effects at their source. Variability in the rating process is not modeled and accounted for in subsequent NAEP analyses.

NAEP analyses model item response probabilities in terms of 1) student proficiency and 2) item characteristics. For NAEP's open-ended items, however, the probability that a given response will earn a particular score depends not only on proficiency and item characteristics, but also on characteristics (e.g., severity) of the person who rates the student's response. This suggests that rater effects should be modeled at the item response level. Item response models for rater and rater-by-item effects—principally variations on the Linear Logistic Test Model (LLTM; Fischer, 1973, 1983)—have been proposed and applied to data arising from performance assessments and other forms of judged performance (e.g., Engelhard, 1994; Wilson and Wang, 1995). LLTMs are generalizations of the one-parameter logistic or Rasch (1960) model, and are more restrictive than NAEP's IRT models. NAEP item responses have been modeled by 2- and 3-parameter logistic

(2PL and 3PL) models and by generalized partial credit (GPC; Muraki, 1992) models, which have not incorporated any rater modeling. Recent advances in statistical model-fitting technology using Markov chain Monte Carlo (MCMC) make it possible to truly generalize the 2PL and GPC models used by NAEP for open-ended items, incorporating rater effects and rater-by-item effects (Patz and Junker, 1997b).

Recent advances in imaging and scoring technology provide us with much more flexibility in the process of distributing open-ended responses to raters. When digitized images of student responses are distributed to raters in a computer network, the possibilities for monitoring rater judgments and providing feedback in real-time are greatly improved over those available using paper-and-pencil technology. Intelligent algorithms that make optimal use of this technology for NAEP are within reach. The effectiveness of such approaches will depend heavily on how well they are adapted to the nature and severity of rater effects in NAEP.

In the present study we begin with a careful, item-by-item analysis of NAEP rater effects, and then explore efficient algorithms for real-time monitoring and feedback for raters. The ultimate goal of this line of research is an elegant simplicity born of careful analysis—a way to increase the reliability of NAEP inferences without adding additional time to NAEP's reporting schedule.

In section 3 we introduce formal notation for IRT models with rater effects within both the LLTM and GLLTM frameworks. In section 4 we review a recent series of studies of rater effects in other IRT contexts in order to place the NAEP challenges in a broader context. We proceed with two analyses of rater effects in two NAEP data sets. Section 5 describes the use of data from the NAEP 1992 Trial State Assessment in Reading at grade 4 in order to 1) to conduct preliminary analyses on a relatively small scale—a convenient extract involving only six items and ten raters was studied, and 2) to carry out a prototype simulation study to investigate the impact of rater effects on item calibration and proficiency estimation under two designs for allocating item responses to raters. Section 6 presents analyses of data from NAEP's 1994 State Assessment in Reading at grade 4. This analysis involved all 22 constructed-response items from the Literary Experience reading scale using the National Comparison Sample. In section 7 we investigate the implications of rater effects for IRT scale scores and classical reliability estimates under three different allocation designs. One of those designs, a stratified randomization based on rater severity, proposes a possible improvement to NAEP's existing randomization design. Finally, in section 8 we draw conclusions and make recommendations for consideration during the redesign of NAEP.

3 IRT models for rater effects

Several studies of rater effects in educational assessment have employed analysis of variance or generalizability methodology in the raw score metric (e.g., Cronbach, Linn, Brennan, and Haertel, 1995; Koretz, Stecher, Klein, and McCaffrey, 1994). When IRT scaling is employed and scale scores reported, as in NAEP, it becomes important to assess the impact of rater variability in the scale score metric. This requires that rater effects be modeled at the item response level. One IRT approach to modeling rater effects is based on the polytomous form of the Linear Logistic Test Model (LLTM; Fischer, 1973, 1983), an extension of the Rasch (1960) model that allows an ANOVA-like additive decomposition in the logit scale. Software to apply restricted cases of the LLTM (so-called *facets* models) has been developed by Linacre (1989), as has software that can estimate models specified under the full LLTM approach (Wu, Adams, and Wilson, in press; Ponocny and Ponocny-Seliger, in press). The technique has been applied to rater effect estimation by Engelhard (1994, 1996), Myford and Mislevy (1995), and Wilson and Wang (1995).

We describe the basic notation for an LLTM IRT rater model here. For J dichotomous items with parameters β_j $(j = 1, 2, \ldots, J)$ presented to I students with proficiencies θ_i $(i = 1, 2, \ldots, I)$ rated by R raters with severity parameters ρ_r $(r = 1, 2, \ldots, R)$, we observe responses $X_{ijr} = x_{ijr}$. Typically every rater does not rate every response, so we let $\{r : r \sim ij\}$ denote the set of raters who rate examinee i's response to item j. A conditional independence assumption is made asserting independence of ratings given rater parameters ρ, item parameters β, and proficiencies θ:

$$p(X|\theta, \beta, \rho) = \prod_i \prod_j \prod_{\{r:r\sim ij\}} p(X_{ijr}|\theta_i, \beta_j, \rho_r). \qquad (1)$$

The distributions of rated responses $p(X_{ijr}|\theta_i, \beta_j, \rho_r)$ follows a binomial distribution with the probability of a correct response given by

$$p_{ijr} = P(X_{ijr} = 1|\theta_i, \beta_j, \rho_r) = \frac{1}{1 + \exp -(\theta_i - \beta_j - \rho_r)}, \qquad (2)$$

that is,

$$logit(p_{ijr}) = \theta_i - \beta_j - \rho_r.$$

This is an example of an LLTM with two facets: one for items and one for raters. LLTMs define a large class of models that include the Rasch model, Masters' (1982) partial credit model (PCM), as well as several models for rater effects. The model is easily extended to include polytomous responses and additional facets, such as those for content domain, rater-by-item interactions, etc.

(Linacre, 1989). For consistency with other notation for existing NAEP models presented below, we will label the particular LLTM in (2) and its extension to the polytomous case the PC-R model, since it is a partial credit model with rater effects.

LLTMs, and the PC-R model in particular, are not generalizations of the IRT models used by NAEP for open-ended items. LLTMs are more restrictive of test items in that they require that all items have a common slope or discrimination parameter. NAEP's GPC model and its 2PL special case allow different items to have different item characteristic curve slopes α_j.

Patz (1996) and Patz and Junker (1997b) introduce a true generalization (called hereafter GLLTM) of the 2PL and GPC models that incorporates rater parameters directly into these models that NAEP currently uses for its open-ended items. GLLTMs generalize LLTMs in the same way that Muraki's (1992) GPC model generalizes Masters' (1982) PCM—by allowing a multiplicative constant in addition to additive constants in the logit scale. We will denote by GPC-R the particular GLLTM that adds rater effects to the GPC model, in analagous fashion to the PC-R designation above. It is important to note, however, that the additive decomposition used to incorporate rater effects in both the PC-R and GPC-R models is quite general. This decomposition in the logit scale results in what Fischer and Parzer (1991) call "virtual items," and these may be used to model not only rater effects but also other experimental conditions or facets (e.g., Huguenard, Lerch, Junker, Patz, and Kass, 1997).

The GPC-R allows individual raters to affect the location parameter for each item, making some items more difficult and others less difficult. Formally, the model lets ρ_{rj} be the severity parameter for rater r on item j. The resulting IRT model may be expressed in terms of its logit:

$$logit(p(X_{ijr}|\theta_i, \alpha_j, \beta_j, \rho_{rj})) = \alpha_j \theta_i - \beta_j - \rho_{rj}. \tag{3}$$

The GPC-R model has the advantage of modeling raters using a model that is a generalization of NAEP's IRT models, but it has the disadvantage of requiring a slower and more cumbersome model-fitting algorithm based on Markov chain Monte Carlo (MCMC). On the other hand, the PC-R model can be fit quickly using the E-M algorithm, but it uses models that are approximations to the NAEP IRT models. In this study we find that the approximation of the GPC-R with the PC-R is reasonably close and may be useful for real-time assessments of rater severity where MCMC would be too slow to be of use.

4 Lessons from other contexts: Rater effects and what we can do about them

4.1 What do rater effects look like?

Using an item response theory approach, several authors have documented the size and scope of rater effects (Engelhard, 1994, 1996; Myford and Mislevy, 1995; Wilson and Wang, 1995). We will use the last of these to illustrate some typical findings. Wilson and Wang (1995), analyzed results from the 1994 California Learning Assessment System (CLAS) test in the topic area of Mathematics. Concentrating on a special sample of the grade 4 students, there were two types of items used that required ratings: investigations (relatively longer items), and open-ended questions (somewhat shorter items). The particular sample studied involved 49 raters. The severities of these raters and their 95% confidence intervals are shown in Figure 1. The intervals do not all overlap, and the chi-square statistic for testing equal severity is 771.14 with 48 degrees of freedom. Therefore, we conclude that, subject to the existing information, and with standard levels of statistical confidence, the raters were operating with different severities. This is an important finding in the present context because CLAS simply added rater judgments without making any adjustments for rater variation. Note that these differences persist even though there were methods in place, such as rater training and checking procedures, that were designed to ameliorate rater severity differences.

To further illustrate the impact of this disparity in rater severity, consider the following. Figure 2 shows item characteristic curves (ICCs) of the investigation item of Form 3 rated by rater 48 (the least severe rater) and rater 46 (the most severe rater). Comparing these two figures, one can easily note that the ICCs shift toward the right from rater 48 to rater 46. It is thus much more difficult for examinees to obtain higher scores from rater 46 than from rater 48. Figure 3 shows the expected scores of this item rated by these two raters. An examinee with ability 0.0 logits would be predicted to have an expected score of 2.2 from rater 48 and 0.7 from rater 46. An examinee with ability 2.0 logits would have an expected score of 3.7 from rater 48 and 1.7 from rater 46. The maximum difference of expected scores derived from these two raters is about 2 points (when examinees' abilities are located between 0.5 logits and 2.5 logits). Since all of the open-ended items and the investigation items are judged on a 6-point scale, a difference of 2 points is an important bias.

This bias is not one that will always be detected by a comparison of raw ratings. For example,

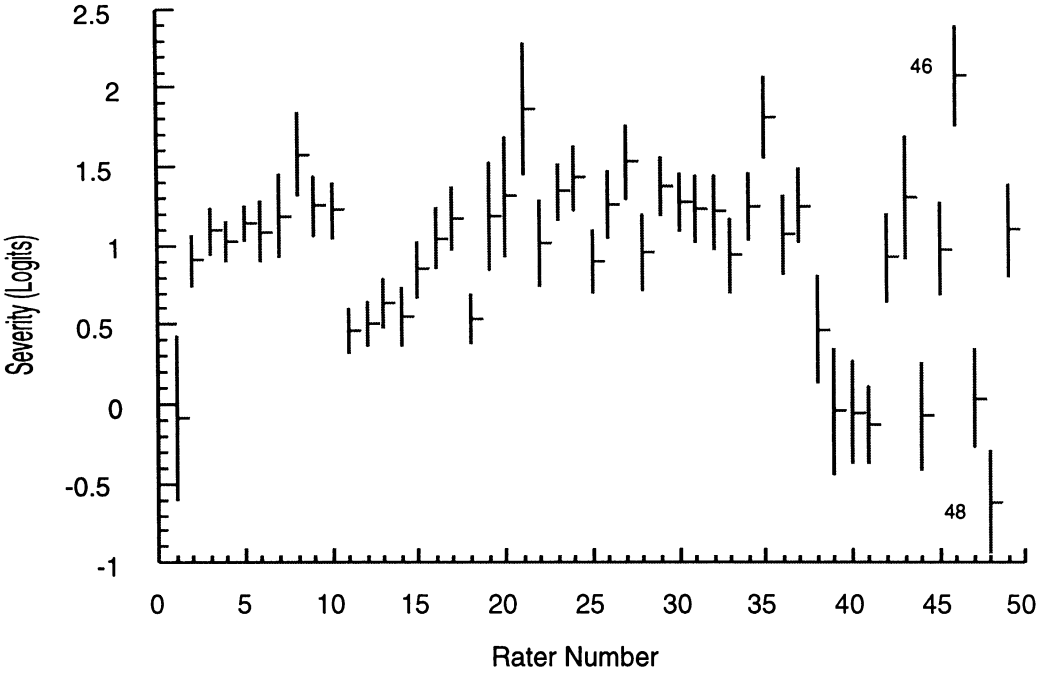


Figure 1: 95% confidence intervals of the 49 rater severities in the CLAS example.

a raw score of 2 derived from rater 48 represents an ability estimate of -0.3 logits, but it would represent 2.4 logits if the score were derived from rater 46. Therefore, in a case where raters vary in severity, the same raw scores derived from two raters are not necessarily the result of the same ability estimates. In other words, raw examinee scores are no longer sufficient statistics for ability estimates (as in the simple logistic model), hence checks on the consistency of raw scores, which have been used as the basis for the traditional measurements of an "industry standard" are not a guarantee against significant problems in rater consistency.

One way that we can examine the effect of variations in rater severity on the results is as follows. Defining severe raters and lenient raters as those whose severities are located one standard deviation (0.56 logits) above and below the mean, respectively, there are 4 severe raters and 7 lenient raters. Suppose the 49 raters are randomly allocated to student scripts, then the probability that an examinee will be judged on an investigation item by a severe rater is $4/49 = 8.2\%$, and by a lenient rater is $7/49 = 14.3\%$. Similarly, the probability that an examinee will be judged on an open-ended item by two severe raters is 0.7%, and by two lenient raters is 2.0%. Fortunately, these percentages are small. If the percentages were larger, then it would call into question the fairness

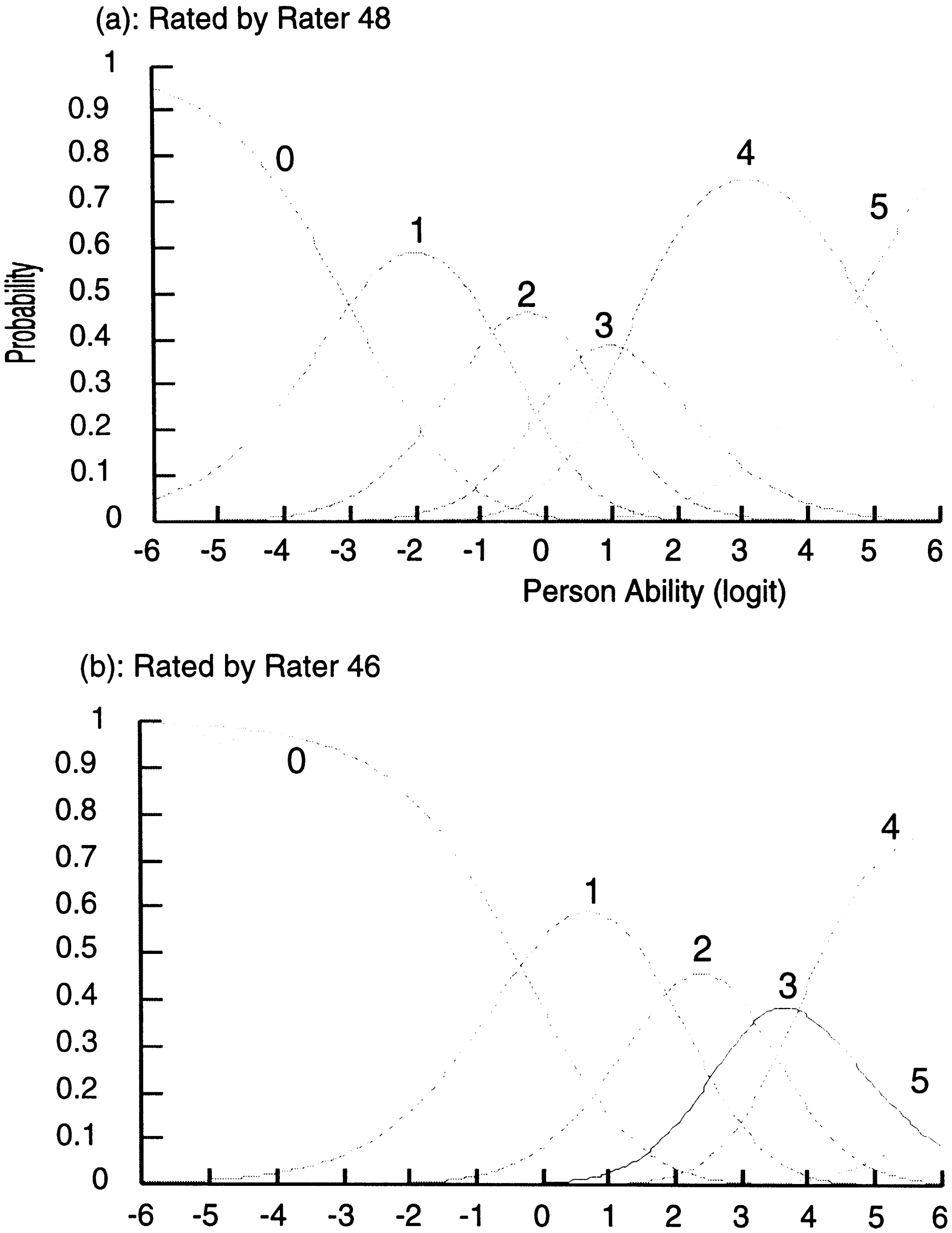


Figure 2: Probability distribution of the investigation item of Form 3 judged by rater 48 (above) and rater 46 (below).

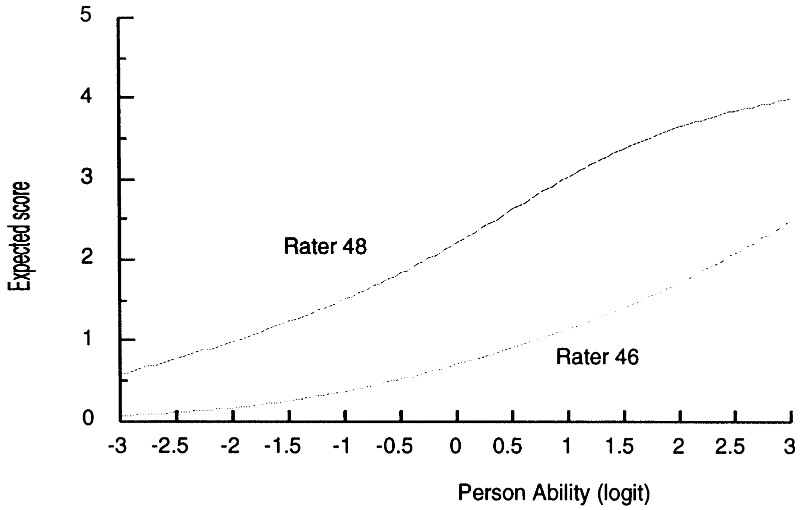


Figure 3: Expected scores on the investigation of Form 3 when the examinees were judged by raters 48 and 46.

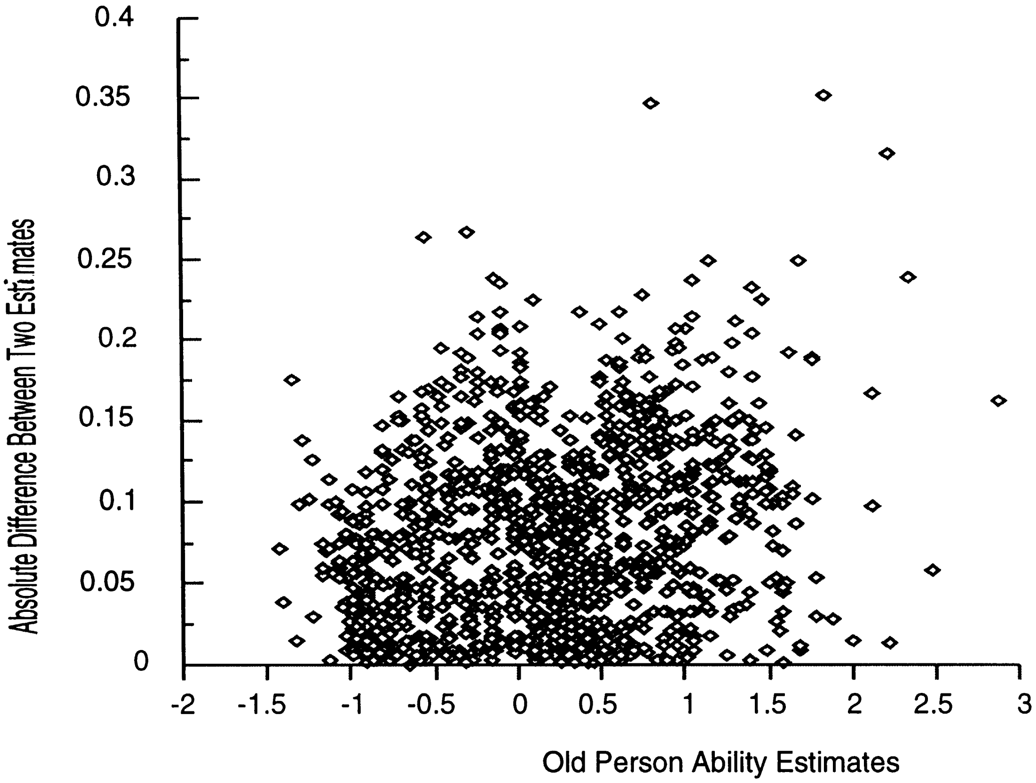


Figure 4: Absolute differences in ability estimates with and without equal rater severity assumption.

of the system as a whole.

Another way to investigate the impact of rater severity on this particular data set is to constrain all of the rater severities to be identical (assuming raters are equal in severity) and then estimate the person ability again. These new estimates are compared to the old estimates where different rater severities are taken into account. We find that the mean of the absolute differences in person ability estimates between these two models is 0.08 logits, and the maximum difference is 0.35 logits. The standard deviation of the estimated absolute differences is 0.06 logits. Figure 4 shows the absolute differences as a function of the old ability estimates.

The influence of variations in rater severities in this particular data is not very great on the test as a whole, because only a few raters differ in severity and because these extreme raters judged mainly the investigation items. This concentration of the consistency problem in the investigation mode may be due to the lack of a second rating (which was used as a quality control method in the open-ended items) for the investigation items. But the differences in rater severities can have large effects on individual students.

Percentages of the examinees	Differences (in logits)	Z-score	Percentiles
Maximum changes	0.35	0.45	17.36
Median changes	0.08	0.10	3.98
75%	0.12	0.15	5.96
90%	0.15	0.20	7.93
95%	0.18	0.23	9.10

Table 1: Changes in percentiles of the person estimates when the raters are assumed to have equal severities.

Assuming a normal distribution of the ability estimates, we derive a rough index of the changes in estimated observed score percentiles when the raters are assumed to have equal severities and show it in Table 1. As the variance of the old ability estimates is 0.61, a maximum absolute difference of 0.35 logits corresponds to a Z-score of 0.45, which in turn corresponds to a change in percentiles of about 17, assuming this person's original position is located at about the mean. (If it is further from the mean, the change in percentiles will be less.) Similarly, the changes in percentiles are below 4 for half of the examinees, below 6 for about 75% of the examinees, and below 8 for about 90% of the examinees. However, the changes in percentiles for about 5% of the examines will be more than 9.

These effects have been found in data that was considered quite acceptable by the standard criterion used by CLAS—the percentage of exact matches. For this particular data set, the percentage of exact matches was 87.5% (CTB/McGraw-Hill, 1995, Table D10), which was within the tolerances set by CLAS, and also quite close to the criterion used by NAEP. Thus, one important message from these findings is that the current practices based on raw score comparisons are not giving us sufficient information to judge whether the raters have been doing a good job.

4.2 Are raters consistent over time?

The example above discussed the dimensions and effects of between-rater differences in severity. Strict interpretation of these results would assume that raters are consistent over the rating period, that is, that within-rater variation was small or nonexistent. An opportunity arose to investigate this in a later rating context in California, again with the CLAS Mathematics Test (Wilson and Case, 1996). On this occasion, the time period during which the ratings took place—morning or afternoon—was recorded. The rating session stretched over 2 1/2 days, so there were five rating periods available for analysis. It was found that raters varied in just about all the ways you could

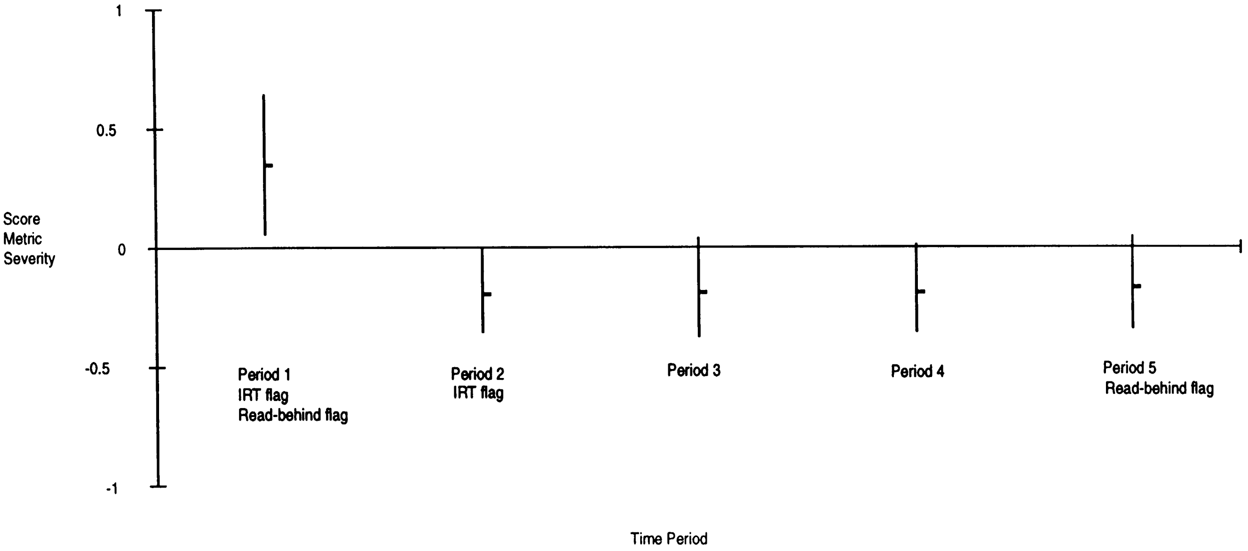


Figure 5: Estimated severity of a CLAS Mathematics rater (32) over five rating periods (Wilson and Case, 1996).

imagine they might. Two examples are shown in Figures 5 and 6. In Figure 5, a rater has started out with an average leniency of almost 40% in score points. This can be translated as meaning that, on average, the rater was assigning four scores out of ten that were 1 score point too high. After the first period, the rater moved back towards the mean over all raters, and in fact became a bit too severe—this sort of over-correction is not unusual. However, this severity was not large enough to reach statistical significance in any of the remaining periods, although it remained constant at about 20% (i.e., on average, the rater was assigning two scores out of ten that were 1 score point too low). Of course, statistical significance may not be the only issue to consider here—a discrepancy of 2 score points out of every 10 on the observed ratings seems fairly large. In Figure 6, the rater has done the opposite—started off pretty much in line with the mean of the raters, then drifted away to become more severe in the last few periods.

The rater severities were of a similar magnitude in this study as the previous one. In order to give some sort of overall indication of the impact of these rater effects, we estimated the average difference between the observed score and the estimated score for three different models: (a) no

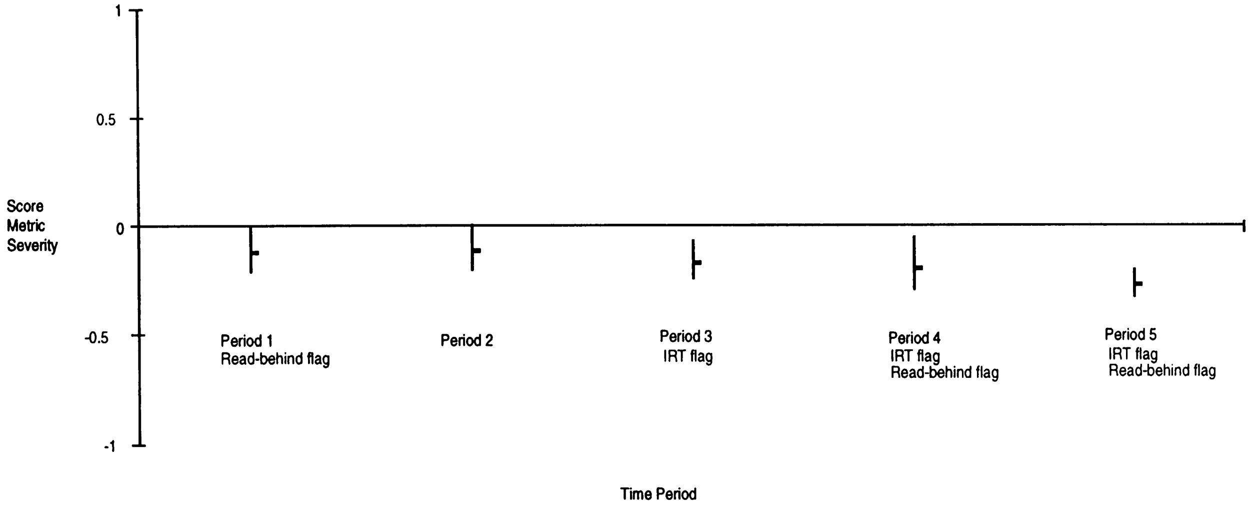


Figure 6: Estimated severity of a CLAS Mathematics rater (65) over five rating periods (Wilson and Case, 1996).

Period	No Rater	Constant Rater	Rater within Period
1	10	7	3
2	12	9	4
3	14	9	5
4	9	7	3
5	12	8	4

Table 2: Impact of rater severities (in percentages) across scoring periods.

rater effects, (b) constant rater effects, and (c) rater effects within period. We calculated these within each period, to see if the results were stable over time. These are shown in Table 2. As can be seen, the estimated reduction in error by introducing constant rater effects is between 2 and 4 percentage points (i.e, on average, the scores would become 2 to 4 points out of 100 more accurate if we consider the raters as having constant severities). This improvement was approximately doubled by considering the raters as having severities that varied between periods.

4.2.1 What can we do to reduce rater variation (both within and between raters)?

Between-rater variation arises initially due to background and personality differences between raters and due to differential effects of training. Ensuring greater uniformity as raters emerge from training would certainly be a positive contribution, but, as has been shown above, raters still have a tendency to drift. Thus, to reduce rater variation in a comprehensive way (both within and between), we need to develop methods of making corrections in an ongoing way. This was attempted in a third study in California, this time using the Golden State Examination in Economics (Hoskens, Wilson, and Stavisky, 1997). The PC-R model was estimated using a marginal maximum likelihood (MML) program (ConQuest; Wu, Adams, and Wilson, in press). Feedback on rater severities (as well as some other basic information) was given to the leaders of small groups of raters (so called "table leaders"). This information was provided after the end of each rating period (approximately a half-day). The overall pattern of rater severities were similar to those described above, so it will not be described here. One way to examine the outcomes of the feedback is to consider the severities in the first and last periods—if the feedback is having a positive effect, then there should be a reduction. Table 3 shows this information. The entries in the cells show how many raters had nonsignificant severities during both periods (top left), significant severities during both periods (bottom right), changed from significant to nonsignificant (bottom left), and changed from

	Final Period			
	All Tables		Four Tables	
Initial Period	nonsignificant	significant	nonsignificant	significant
nonsignificant	14	2	14	2
significant	7	5	6	1

Table 3: Number of raters that were, or were not, significantly different from the average for initial and final rating periods.

nonsignificant to significant (top right). First this is shown for all raters (left-hand panel in Table 3). The good news is that seven raters have reduced their severities from significant to nonsignificant. The not-so-good news is that five have maintained their severities as significant, and two have actually increased their severities so that they have become significant. There was one complication at this scoring site: One of the table leaders became very opposed to the prevailing standards that were being applied to the students' work. He advocated considerably "higher" standards (i.e., increased severity), and his table was greatly affected by this conflict, with raters changing their severities quite dramatically during the scoring session (in the end, this table leader left the scoring session before the beginning of period 5). If we remove the raters who were part of this table, then the results are shown in the right-hand panel in Table 3. Here the number of raters who maintained their severities so that they were statistically significant at both the beginning and end has been reduced to 1. The removal of this group of raters has no effect on the number of raters who changed from nonsignificance to significance.

A second strategy to reduce rater effects is to control for them statistically. This can be done by retaining the rater parameters in the statistical model used to scale the data. Effectively, this is what was done in Table 2, and so the interpretations of effect size that were shown there are indicative of the potential overall effects of such adjustments. This is a strategy that has not been pursued much in large-scale assessments. This is partly because the testing agencies have been satisfied with success rates such as those noted above for CLAS: 90% (or so) exact matches using double-readings. As we have shown above, this overall statistic is quite capable of concealing some very large problems, and probably does so in many circumstances. Interestingly enough, this is very close to the same criterion that was used by NAEP to accept the rescored performance assessments in the 1992 data; the rates for the 1994 NAEP data hovered around this figure, some better, some worse. Of course, sensible rater allocation policies (i.e., ensuring that each student's work is scored by several raters) will assuage the effects of bias on individual student results. And, in a case such as NAEP, where group rather than individual results are the focus, the effects of having several raters scoring the group's results will also reduce the problem of bias. However, in this case, the effects of rater inconsistency will be propagated to the final results in the form of underestimated error variance rather than as bias.

5 NAEP analyses Part I: 1992 Trial State Assessment in Reading

In this section we first describe the rater-by-item effects observed in the 1992 NAEP data set, and then we describe a preliminary simulation study designed to evaluate several designs for distributing responses to raters in light of these effects.

5.1 GLLTM analyses of rater effects

Patz and Junker (1997b) fit a GLLTM (in particular, the GPC-R model of equation 3) to a subset of the data from NAEP's 1992 Trial State Assessment Program in Reading at grade 4. The subset involved 1,500 students whose responses to six open-ended items were rated by one of the ten most common raters. The purpose of the analysis was to understand the types of rater effects present in data sets of that type and to explore effective ways of modeling them.

Figure 7 depicts the fitted item-rater characteristic curves for the first item and for the set of ten raters. This figure illustrates the manner in which rater effects are being modeled here—each rater has the effect of shifting the curve of each item, which is consistent with the way rater effects are modeled in studies of rater effects and rater feedback described in section 4 above. Figure 7 also communicates the nature and severity of rater effects in terms of raw item score—the probability of obtaining credit for a response may vary by as much as 20% depending on the rater assigned to rate the response. Seen another way, out of 10 average students, the most severe rater would be expected to fail two more students than the least severe rater.

The model was fit using the Metropolis-Hastings within Gibbs algorithm described in Patz (1996) and Patz and Junker (1997).

Figure 8 shows the estimated posterior distributions for rater-by-item effects ρ_{rj} for all ten raters on each of the six items. Rater 6 makes item one "easy" whereas rater 8 makes it more difficult, for example. In Figure 8 one can detect heterogeneity in both the overall (mean) severity of raters and also the differential severity of raters across the items. The variance of the estimated

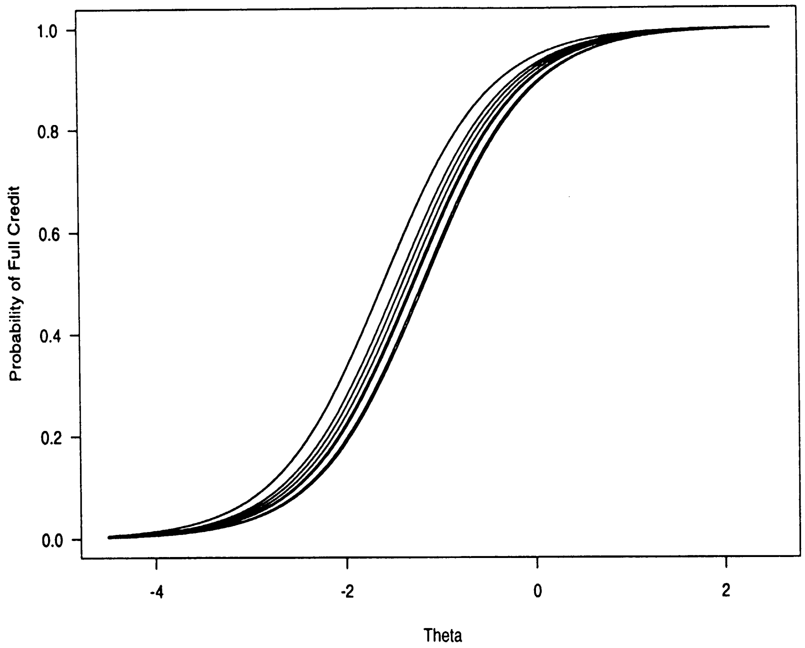


Figure 7: Fitted item-rater characteristic curves for one item and ten raters, from a subset of NAEP's 1992 Trial State Assessment Program in Reading (Patz, 1996).

rater-by-item effects $\hat{\rho}_{rj}$ is 0.112, which means that the standard deviation of these effects is about one third of the theoretical (*a priori*) proficiency distribution. The variance of the mean of the estimated rater-by-item effects for raters across items, $\hat{\rho}_{r\cdot}$, is 0.0545, meaning that about half of the variance of the estimated rater-by-item effects is attributable to a general tendency of raters to be severe or lenient across items.

5.2 Assessing rater designs by simulating from a fitted rater model

Rater effects of the type depicted in Figure 8, when present, are typically ignored in standard analyses of item response data involving constructed-response items, except for the work using PC-R models discussed above. In this section we investigate the implications that ignoring these effects may have on inferences regarding item parameters and student proficiencies.

Table 4 compares posterior means and standard deviations from an MCMC fitting of the standard 2PL model with those of the rater effect model described above. Note that the estimated slope parameters, α_j, remain largely unchanged, whereas there are some significant changes in the location parameters, β_j.

Table 4 raises an important question about the implications of ignoring systematic rater effects

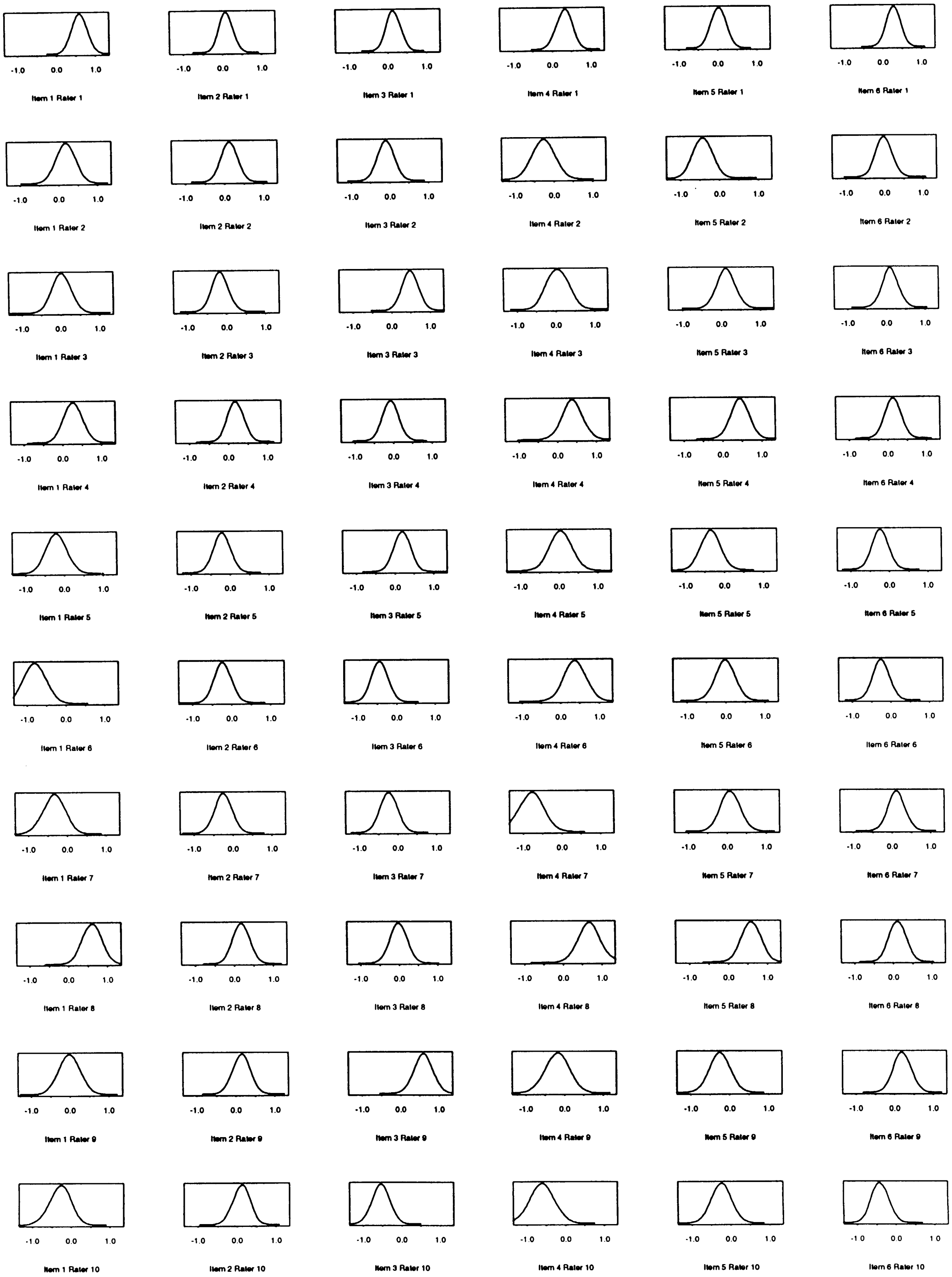


Figure 8: Posterior distributions for rater-by-item effects ρ_{rj} from a subset of 1992 NAEP data involving 6 items and 10 raters.

Param.	2PL Model	Rater Model
β_1	-1.73 (0.10)	-1.19 (0.19)
β_2	-0.45 (0.07)	-0.37 (0.16)
β_3	0.42 (0.07)	0.57 (0.17)
β_4	-1.54 (0.12)	-1.32 (0.21)
β_5	-0.99 (0.09)	-0.98 (0.18)
β_6	0.26 (0.07)	0.54 (0.16)
α_1	1.40 (0.13)	1.39 (0.14)
α_2	1.08 (0.10)	1.09 (0.10)
α_3	1.17 (0.11)	1.19 (0.11)
α_4	2.09 (0.20)	2.19 (0.22)
α_5	1.60 (0.14)	1.64 (0.14)
α_6	1.11 (0.11)	1.11 (0.10)

Table 4: MCMC parameter estimates for $J = 6$ 2PL items based on a sample of $I = 1,000$ students whose responses were rated by one of $R = 10$ raters in NAEP's 1992 Trial State Assessment in Reading.

when they are present. We address this question using a straightforward simulation study.

In this simulation we varied two conditions:

- **Rater effect type** was classified in one of four conditions depending on the overall variance of rater-by-item effects and on the proportion of that variance attributable to overall severity/leniency of individual raters across items. The first condition represents a control condition—no rater effects are present, and data is generated from a standard 2PL model. The second rater effect condition reproduces the nature of the rater effects observed in the NAEP subset and depicted in Figure 8. Here the standard deviation, $\sigma_{\rho_{rj}}$, of the rater-by-item effects was 0.33, and 54% of the variance is attributable to overall (mean) rater effects across items (i.e., $\sigma^2_{\bar{\rho}_{r\cdot}} = 0.54\sigma^2_{\rho_{rj}}$). The third and fourth conditions represent a somewhat more serious rater-by-item variability ($\sigma_{\rho_{rj}} = 0.66$), but they differ in the proportion of variance attributable to mean rater effects: In the third condition raters vary primarily in terms of overall severity, whereas in the fourth condition rater variability is heterogeneous across items.

- **Rater-to-task design** had two conditions. In 1992 NAEP, raters were randomly assigned to student papers, but one rater scored all performances by the student. This assignment "by student" is the first rater-to-task condition. In the second condition, raters are randomly assigned to student responses: all raters rate some responses to all items, but each response by each student is rated by a randomly selected rater. This is the "random" condition for raters-to-task design.

	Effect	None	Mild	Mod. Overall	Mod. by Item
	std. dev. of ρ_{rj}	0.00	0.33	0.66	0.66
Design	% of var. in $\rho_{r\cdot}$	—	0.54	0.90	0.10
By Stdnt	Locations β_{1j}	0.075 (0.002)	0.102 (0.003)	0.139 (0.027)	0.206 (0.004)
	Slopes β_{2j}	0.101 (0.003)	0.107 (0.003)	0.127 (0.005)	0.164 (0.003)
	Proficiencies θ	0.582 (0.004)	0.597 (0.004)	0.708 (0.013)	0.594 (0.004)
Random	Locations β_{1j}	0.075 (0.002)	0.121 (0.017)	0.133 (0.014)	0.224 (0.019)
	Slopes β_{2j}	0.101 (0.003)	0.123 (0.005)	0.168 (0.005)	0.173 (0.007)
	Proficiencies θ	0.582 (0.004)	0.601 (0.010)	0.612 (0.010)	0.614 (0.014)

Table 5: Mean (across 100 simulated data sets) of the RMSE for item parameters and proficiency estimates. Standard errors of these means are in parentheses.

We simulated data sets with performances by 2,000 examinees on 12 two-level constructed-response items. For each experimental condition, 100 data sets were generated. First, student proficiencies were generated according to a $N(0,1)$ distribution. Then item parameters α and β were generated in a manner consistent with observed distributions of the estimated parameters in NAEP's 1992 Trial State Assessment Program in Reading at grade 4. In particular, α_j's were generated according to a log-normal$(0.34, \sigma = 0.24)$ distribution, and β_j's were generated according to a $N(-0.13, \sigma = 1.19)$ distribution. Rater effect parameters, ρ_{rj}, for the ten raters on twelve items, were not generated randomly but were held fixed at equally spaced quantiles of the normal distributions implied by their experimental condition.

Each generated data set was fit to the standard 2PL model using an E-M-based marginal maximum likelihood IRT model-fitting software package (PARDUX; Burket, 1996). For each data set, the square root of the mean squared error (RMSE) was calculated for the twelve α's, the twelve β's, and the 2,000 θ's. The mean (across the 100 simulations) of these RMSE statistics are presented in Table 5, along with their associated standard errors.

5.2.1 Discussion of the first simulation study

The results of this simulation, which are presented in Table 5, suggest several conclusions. First, rater effects of this type, when present but not modeled, increase the error in the estimation of item parameters. This increase is most notable in the location parameters, β_j, and this increase is not sensitive to the design for assigning raters to responses within examinee, at least, among the designs investigated here. Even the fairly mild rater effects observed in the NAEP example increase the error in item location estimation by about one third. Estimation of the slope parameters, α_j, is not seriously affected in this case.

The impact of not-modeled rater effects on proficiency estimation is considerable when these effects are systematic within rater (i.e., of the 'overall' variety) and when the same rater scores all responses by a given examinee. Not surprisingly, the impact of these effects is significantly mitigated when an individual examinee's responses are rated by a random selection of raters. Since it is difficult to know *a priori* the nature and severity of rater effects that may be encountered in scoring examinees, it appears wise to randomize the assignment of individual item responses to raters whenever possible.

This example from NAEP demonstrates that rater effects can be incorporated into the NAEP item calibration model yielding useful information about the characteristics of raters and items, a finding that is entirely in agreement with the results of the earlier series of studies cited above.

In the context of the present study, we can conclude that

1. The simulation methodology is workable and yields useful information regarding the distribution design for assigning responses to raters.

2. Measurement quality could be significantly improved by randomly assigning raters to item responses, instead of assigning raters to examinees and having just one rater scoring all responses by the examinees. In 1994 NAEP implemented this change, and information from the simulation study suggests that this change was a significant improvement and that it should be preserved in the redesign.

5.3 An LLTM analysis of rater effects in 1992 NAEP

Although the results in section 5.2 show that fitting the GPC-R models can make important improvements, there is a serious limitation to that usefulness—the MCMC estimation is very slow. For practical purposes of providing real-time feedback to raters about their performances, MCMC fitting of GLLTMs is too slow to be useful. Thus we would like to know whether the faster MML estimation technique, applied to the PC-R model, would supply useful information.

We fit three LLTMs to this 1992 NAEP extract:

1. A regular partial credit model (PCM) that ignores potential rater effects and estimates only item difficulties and steps (based on all ratings available for each item).

	Model	-2logL	No. Param	AIC
NAEP 92	1. Partial credit	10526.4	7	140.4
	2. General rater effects	10491.0	16	123.0
	3. Item-specific rater effects	10406.5	61	128.5

Table 6: Goodness of fit of three LLTM models for the NAEP 1992 data extract.

	Item 1	Item 2	Item 3	Item 4	Item 5	Item 6	Mean
PC-R	0.243	0.351	0.357	0.278	0.312	0.433	0.329
GPC-R	0.238	0.379	0.367	0.203	0.281	0.387	0.309

Table 7: Mean absolute residuals, $|X_{ij} - P_{ij}(\theta_i)|$, resulting from fitting LLTM (PC-R) and GLLTM (GPC-R) models to the 1992 NAEP data extract.

2. A PC-R model with general rater effects that includes parameters for rater severity that are constant over items, in addition to item difficulties and item step parameters.

3. A PC-R model with item-specific rater effects that includes rater severity parameters that are specific to each item, in addition to item difficulties and item steps. The item-specific rater parameters indicate how much more severe (or lenient) a rater is than the average rater when scoring a particular item.

Table 6 presents goodness of fit results from fitting the three models. Likelihood-ratio test statistics indicate that the model goodness of fit to the data significantly improves when general rater effects are taken into account in addition to item difficulties and steps ($\chi^2_9 = 35.4$, $p < 0.01$), and that further improvement is obtained when the rater effects are modeled to be item specific rather than general ($\chi^2_{45} = 184.5$, $p < 0.01$). From the AIC indices, however, one could conclude that the model with general rater effects fits the data best.

5.4 LLTM vs. GLLTM

Table 7 compares the residuals obtained fitting the PC-R and GPC-R models to the same extract of data from 1992 NAEP. Overall, the mean residual is lower for the GPC-R model, although this varies by item. Figure 9 compares estimated rater-by-item effects resulting from an PC-R analysis of the 1992 data set with those obtained using the MCMC fit of the GPC-R.

These results suggest that the more efficiently estimated PC-R model may provide a useful real-time approximation to the GPC-R rater severity. Such a real-time estimate of rater severity may be useful in providing feedback and modifying allocation strategies, as discussed in section 7 below. The similarity is also displayed in the bottom panel of Figure 11, which shows ICCs for a partial credit model and several different estimates of a GPC model.

6 NAEP analyses Part II: 1994 State Assessment in Reading

Using the preliminary 1992 data analyses as a guide, we conducted a second set of analyses. For this study we used all of the constructed-response data on the "Reading for Literary Experience" scale from the National Comparison Sample in NAEP's 1994 State Assessment Program in Reading. In particular, the data set has $N = 4,610$ examinees; $J = 22$ items (fourteen 2-level constructed-response items, four 3-level constructed-response items, and four 4-level constructed-response items); $R = 64$ raters; and second ratings on 25% of the items. A large portion of this data set is missing by design, according to NAEP's matrix sampling design.

6.1 Calibrations: NAEP, MCMC, PCM

We began our model-fitting analysis by fitting NAEP's item response theory (IRT) models to our particular data extract using both the MCMC model-fitting technology and marginal maximum likelihood model-fitting technology. This exercise serves to 1) verify the plausibility of the MCMC parameter estimates vis-a-vis those reported by NAEP and 2) provide information about the speed of the model-fitting algorithms with the current data set.

For open-ended items, NAEP uses Muraki's (1992) generalized partial credit (GPC) model. We present the model here in a slightly different (but equivalent) parameterization than that used by NAEP in its technical reports:

$$p(X_{ij} = k|\theta_i, \alpha_j, \beta_{1j}, \beta_{2j}, \ldots, \beta_{Kj}) = \frac{\exp \sum_{l=1}^{k}(\alpha_j\theta_i - \beta_{lj})}{\sum_{\nu=1}^{K} \exp \sum_{l=1}^{\nu}(\alpha_j\theta_i - \beta_{lj})} \tag{4}$$

where X_{ij} is the (rated) response of examinee i to item j, β_{kj} is a category k ($k = 1, 2, \ldots, K$) location parameter for item j ($\beta_{1j} \equiv 0$), and α_j is a slope or discrimination parameter for item j.

The parameter estimates we obtain from an MCMC fit are not expected to be identical to those reported by NAEP for several reasons, most notably that we are using a different data set, but also due to slight differences in parameterizations that lead us to specify slightly different prior

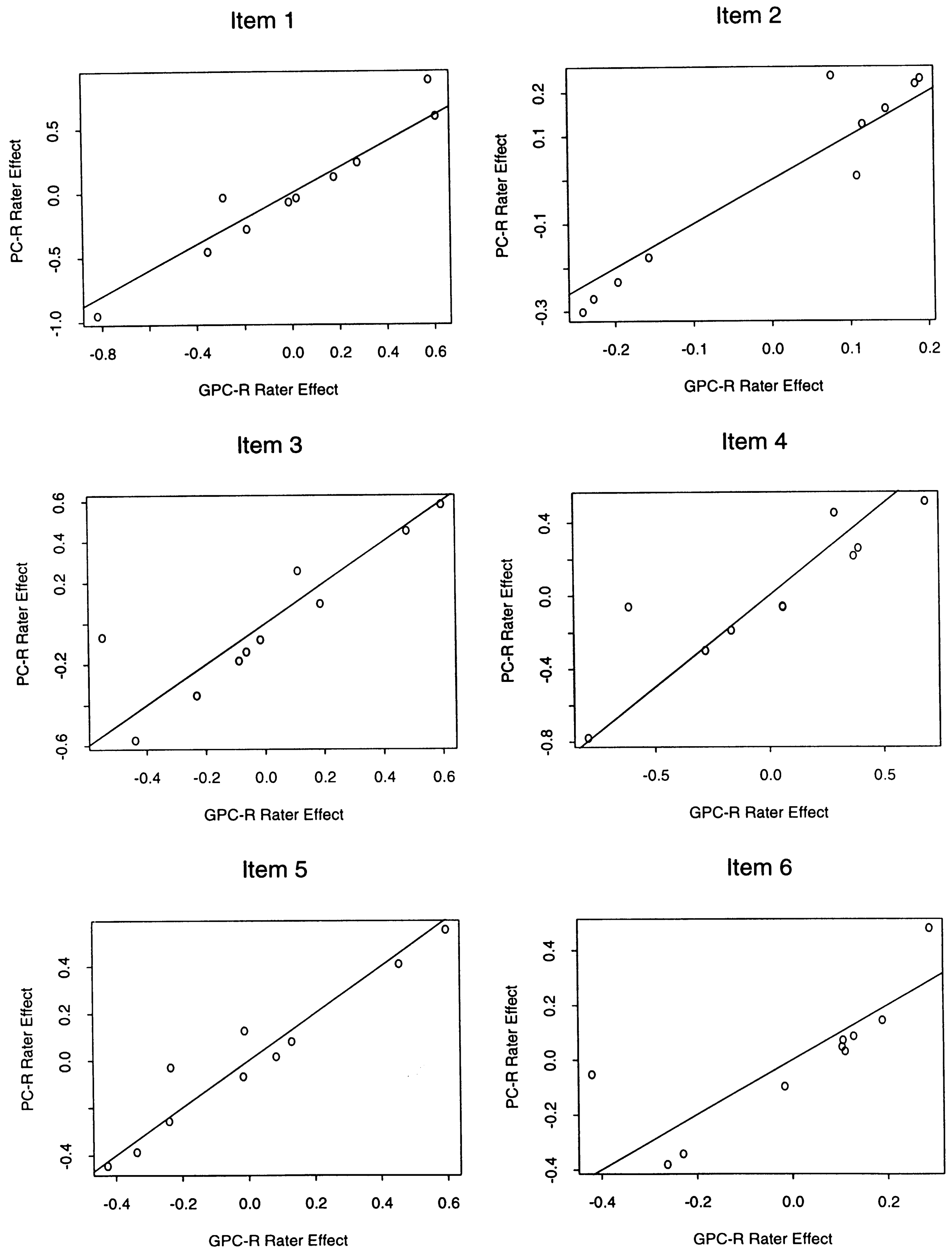


Figure 9: Comparison of rater-by-item effects estimated using MCMC fitting of the GLLTM (GPC-R model) and the faster E-M fitting of the LLTM (PC-R model). Rankings based on severity are reasonably similar between these two approaches.

distributions. Patz and Junker (1997b) report very precise agreement between 2PL parameter estimates obtained through MCMC and those obtained using BILOG on the same data set.

The MCMC parameter estimates presented in Table 8 are based on a run of $10,000$ iterations of a Markov chain following a "burn-in" of 1,000 iterations. The maximum Monte Carlo standard error associated with these estimates is 0.05, suggesting that small differences between these MCMC estimates and those from MML and NAEP should not be over-interpreted at this point. Although greater precision in MCMC estimates may be obtained from longer runs of the Markov chain, this seemed unnecessary for our purposes here.

The information in Table 8 is depicted graphically in Figure 10. We can see that location parameters are generally very close, especially between MCMC and MML. Slope parameter estimates for MCMC are systematically smaller than those reported by NAEP and those fit under MML. This warranted some further investigation, especially with respect to the impact of the prior distributions on these parameters. Further investigation revealed that more diffuse priors had only minimal impact on estimated parameters.

6.2 Unbalanced allocation designs

Of primary importance for the present study is the distribution of item responses to the set of raters. Figure 12 depicts a table showing the number of responses to each item that are rated by each rater.

The design in the assignment of raters to items has implications for our ability to detect and correct any rater effects. This is a general issue that holds for IRT analyses but for other methodologies as well, such as generalizability theory.

Consider, for example, a situation where various raters rate partially overlapping sets of items, as is the case for each of the item clusters shown in Figure 12. Such a situation precludes us from investigating the generality of rater effects over items, as estimates of rater main effects will be confounded with differences in difficulty of the items that the various raters rated. Similarly, estimates of item difficulty will be confounded with differences in severity between groups of raters. Consider, in particular, the fourth cluster of items (items 17 through 22) that is displayed in the most right-hand panel of Figure 12. Two major groups of raters can be distinguished, those that rate the first three items of the cluster (raters 452 through 457) , and those that rate the first two and the last three (raters 467 through 479). Suppose that both groups of raters have the same

NAEP ID	Param	MCMC Fit	MML Fit	NAEP Rept
R012002	α	1.82	2.146	2.324
R012004	α	1.17	1.342	1.278
R012008	α	0.95	1.084	0.967
R012010	α	1.90	1.999	1.812
R012102	α	1.10	1.184	1.125
R012104	α	1.02	1.112	1.120
R012106	α	1.36	1.687	1.503
R012108	α	1.15	1.351	1.124
R012109	α	0.97	1.056	0.865
R012112	α	1.11	1.299	1.248
R012601	α	0.94	1.141	1.467
R012604	α	1.24	1.524	2.006
R012611	α	0.93	1.306	1.353
R015802	α	0.66	0.752	0.688
R012002	β	-0.28	-0.306	-0.432
R012004	β	0.52	0.525	0.431
R012008	β	-0.55	-0.561	-0.569
R012010	β	-0.67	-0.651	-0.794
R012102	β	0.01	0.006	-0.115
R012104	β	-0.27	-0.276	-0.397
R012106	β	0.04	0.060	0.081
R012108	β	-1.21	-1.247	-0.234
R012109	β	-1.26	-1.253	-0.514
R012112	β	-0.98	-0.950	-1.027
R012601	β	1.37	1.433	1.687
R012604	β	1.80	1.914	2.112
R012611	β	0.19	0.246	0.290
R015802	β	-0.72	-0.710	-0.878
R015803	α	0.83	1.014	1.010
R015806	α	0.96	1.083	1.049
R015807	α	0.84	1.015	0.994
R015808	α	0.81	0.956	0.986
R015803	β_1	-1.52	-1.573	-1.834
R015803	β_2	1.44	1.524	1.557
R015806	β_1	-0.91	-0.900	-1.069
R015806	β_2	1.64	1.712	1.820
R015807	β_1	-1.03	-1.063	-1.325
R015807	β_2	0.98	1.056	1.002
R015808	β_1	-1.19	-1.189	-1.398
R015808	β_2	1.55	1.644	1.432
R012006	α	0.36	0.788	0.819
R012607	α	0.51	1.161	1.530
R015804	α	0.48	1.510	0.971
R012006	β_1	0.55	1.270	1.522
R012006	β_2	0.75	0.653	0.352
R012006	β_3	0.47	0.626	0.675
R012607	β_1	0.36	0.808	0.813
R012607	β_2	-0.28	-0.557	0.900
R012607	β_3	-0.09	0.16	2.018
R015804	β_1	2.71	3.472	2.930
R015804	β_2	0.87	0.926	-1.659
R015804	β_3	1.10	2.099	1.065
R012111	α	1.77	3.391	NA
R012111	β_1	-1.07	-1.464	NA
R012111	β_2	0.90	1.242	NA
R012111	β_3	2.53	3.428	NA

Table 8: Parameter estimates from MML and MCMC fits of open-ended items from the literary experience scale of the NAEP 1994 State Assessment in Reading. Estimates should be similar but not identical. MML and MCMC fits come from National Comparison Sample only, MCMC are expected a posteriori (EAP) estimates using fairly disperse prior distributions. MML estimates were obtained using PARDUX. NAEP collapsed levels on item R012111; our MCMC and MML analyses did not.

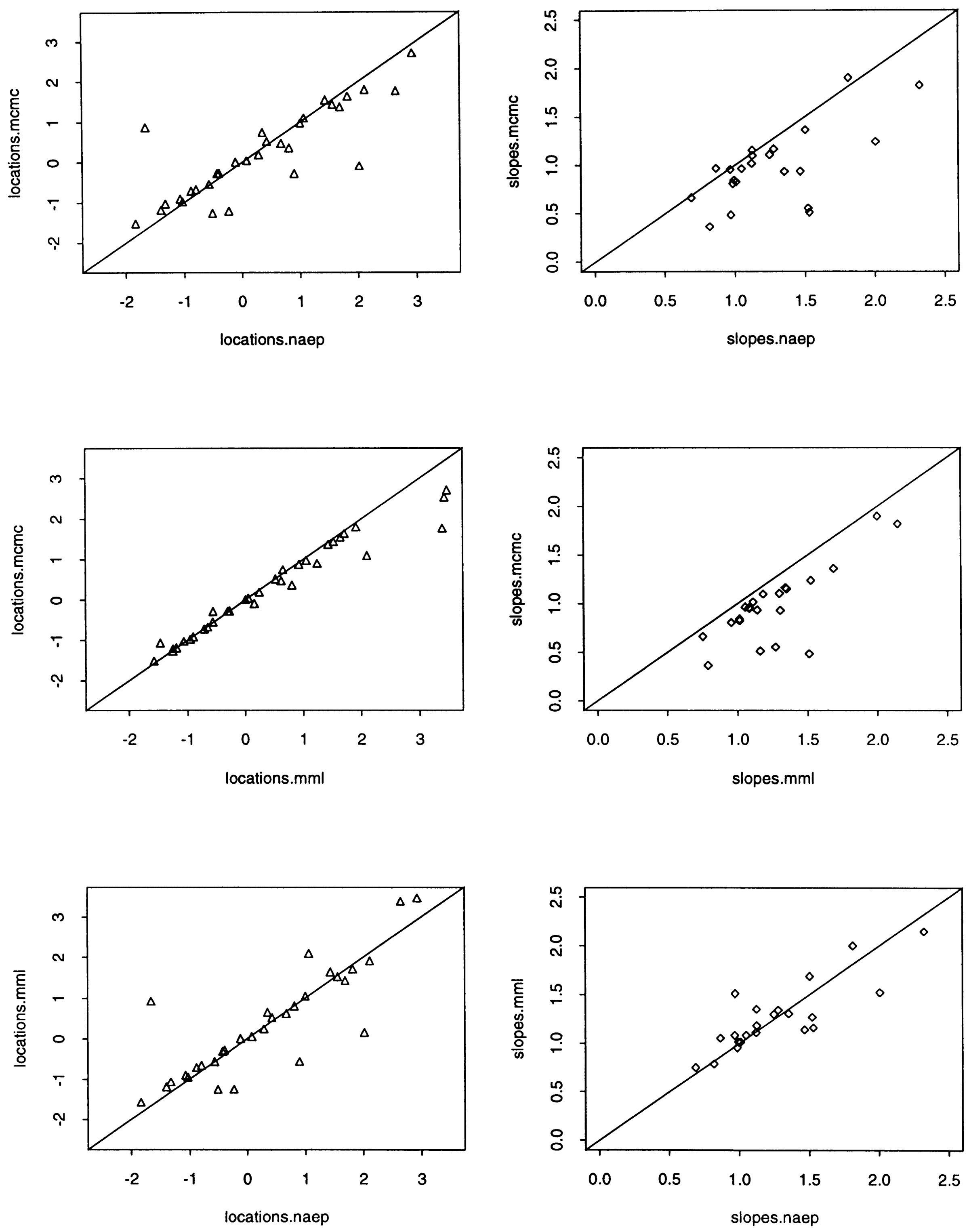
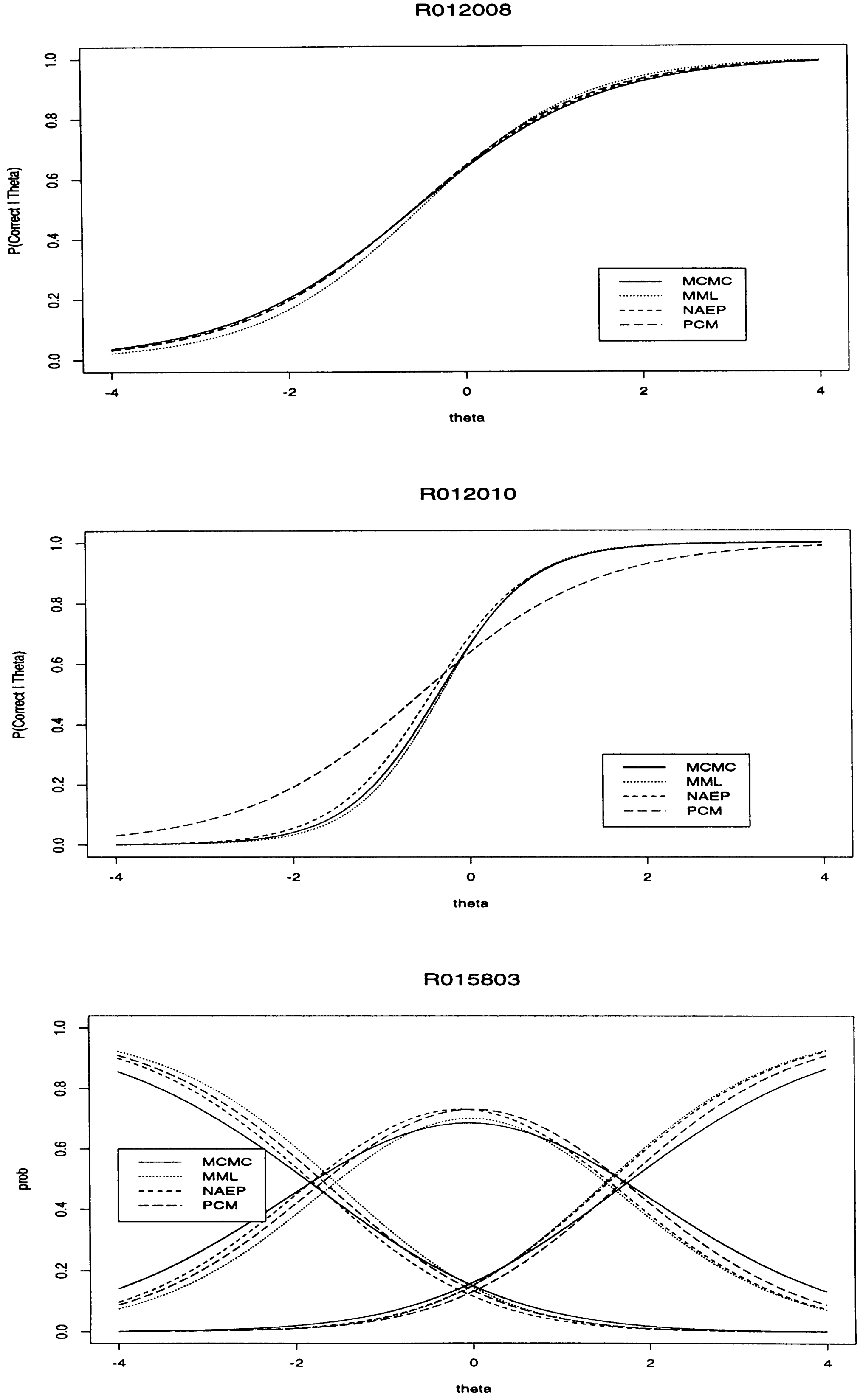


Figure 10: Comparison of MCMC parameter estimates with those reported by NAEP and those obtained using an MML algorithm. Location parameters (β) are on the left, and discrimination (or slope) parameters (α) are on the right.



R012008

R012010

R015803

Figure 11: Comparison of fitted ICCs from MCMC, MML, NAEP (reported), and PCM for three open-ended items. The fitted ICCs are generally quite close, differing most noticeably for two-level items with relatively high or low estimated (GPC) discrimination parameters α.

distribution of severity (for illustrative purposes), but that the items increase in difficulty going from the first item to the last one in the cluster. Then, in contradiction to the initial assumption, the second group of raters will appear to be more severe than the first group, only because of the way that raters were assigned to student responses in the NAEP design. For this reason, we do not fit a "main effects" rater model to this 1994 NAEP extract, whereas we were able to fit such a model to the (artificially) balanced 1992 extract in section 5.3 above.

Also problematic is the uneven number of ratings provided by individual raters. Item 1, for example, was rated 14 times by rater 310 and 440 times by rater 311. Estimation of rater severity for items with very few ratings is problematic, and this type of unbalance also complicates interpretation of estimated rater severity parameters.

An optimal situation for monitoring the impact of rater effects is one where the design in the assignment of raters to items is balanced, where the two facets, raters and items, are completely crossed. In such a case, problems like the one described above would then be avoided. A completely crossed design may not be feasible, given logistical constraints involved in NAEP. Nonetheless, an appropriate partially balanced design, intended to facilitate the detection of rater-by-item bias, would be a significant improvement.

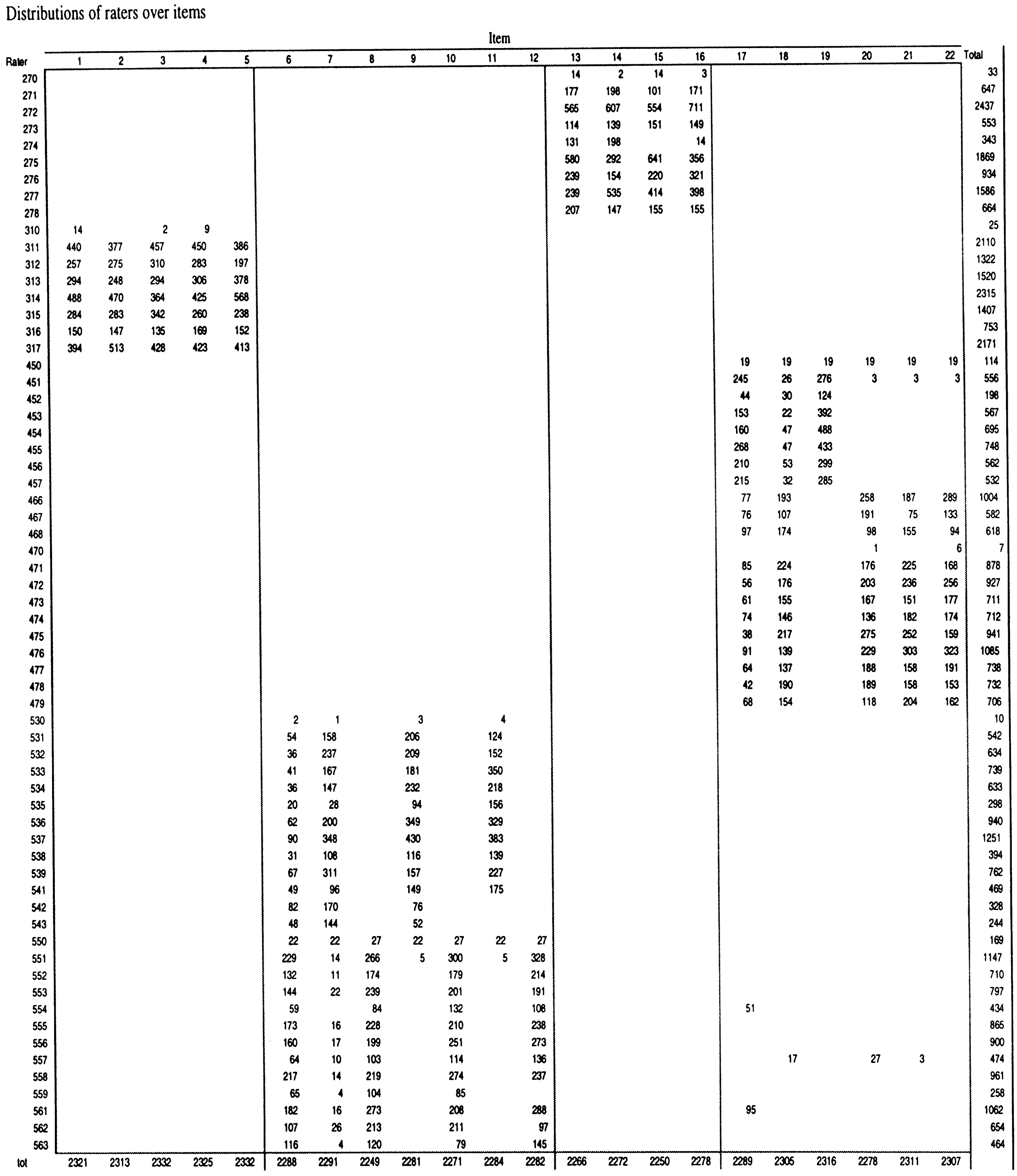
6.3 LLTM analysis of rater effects in 1994 NAEP

GLLTMs (and the GPC-R in particular) generalize NAEP's existing IRT models, allowing us to characterize the consequences of rater errors in terms of existing NAEP variables (item parameters, scale scores, etc.). Unfortunately, the technology for fitting the GPC-R is too slow to use for real-time rater diagnosis and feedback purposes.

PC-Rs, however, may be fit much more quickly and may have use in real-time applications. PC-Rs are special cases of the models used by NAEP. In this section we describe two PC-Rs that were fit to the 1994 data extract described in Section 6 above:

1. A regular partial credit model (PCM) that ignores potential rater effects and estimates only item difficulties and steps (based on all ratings available for each item).

2. A PC-R model with item-specific rater effects that includes rater severity parameters that are specific to each item, in addition to item difficulties and item steps. The item-specific rater parameters indicate how much more severe (or lenient) a rater is than the average rater scoring a particular item.



Distributions of raters over items

Item

Rater	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	Total
270													14	2	14	3							33
271													177	198	101	171							647
272													565	607	554	711							2437
273													114	139	151	149							553
274													131	198		14							343
275													580	292	641	356							1869
276													239	154	220	321							934
277													239	535	414	398							1586
278													207	147	155	155							664
310	14		2	9																			25
311	440	377	457	450	386																		2110
312	257	275	310	283	197																		1322
313	294	248	294	306	378																		1520
314	488	470	364	425	568																		2315
315	284	283	342	260	238																		1407
316	150	147	135	169	152																		753
317	394	513	428	423	413																		2171
450																	19	19	19	19	19	19	114
451																	245	26	276	3	3	3	556
452																	44	30	124				198
453																	153	22	392				567
454																	160	47	488				695
455																	268	47	433				748
456																	210	53	299				562
457																	215	32	285				532
466																	77	193		258	187	289	1004
467																	76	107		191	75	133	582
468																	97	174		98	155	94	618
470																				1		6	7
471																	85	224		176	225	168	878
472																	56	176		203	236	256	927
473																	61	155		167	151	177	711
474																	74	146		136	182	174	712
475																	38	217		275	252	159	941
476																	91	139		229	303	323	1085
477																	64	137		188	158	191	738
478																	42	190		189	158	153	732
479																	68	154		118	204	162	706
530						2	1		3		4												10
531						54	158		206		124												542
532						36	237		209		152												634
533						41	167		181		350												739
534						36	147		232		218												633
535						20	28		94		156												298
536						62	200		349		329												940
537						90	348		430		383												1251
538						31	108		116		139												394
539						67	311		157		227												762
541						49	96		149		175												469
542						82	170		76														328
543						48	144		52														244
550						22	22	27	22	27	22	27											169
551						229	14	266	5	300	5	328											1147
552						132	11	174		179		214											710
553						144	22	239		201		191											797
554						59		84		132		108					51						434
555						173	16	228		210		238											865
556						160	17	199		251		273											900
557						64	10	103		114		136						17		27	3		474
558						217	14	219		274		237											961
559						65	4	104		85													258
561						182	16	273		208		288					95						1062
562						107	26	213		211		97											654
563						116	4	120		79		145											464
tot	2321	2313	2332	2325	2332	2288	2291	2249	2281	2271	2284	2282	2266	2272	2250	2278	2289	2305	2316	2278	2311	2307	

Figure 12: Distribution (frequencies) of item responses to raters for the 1994 NAEP State Assessment in Reading National Comparison Sample. The seriously unbalanced distribution complicates analyses of rater-by-item bias.

	Model	-2logL	No. Param	AIC
NAEP 94	1. Partial credit (PCM)	62705.2	35	775.2
	2. Item-specific rater effects (PC-R)	62331.2	291	913.2

Table 9: Goodness of fit for two PC-R models for the NAEP 1992 and 1994 data sets.

We do not fit a "main effects" rater model to this data set for the reasons mentioned above in section 6.2.

Table 9 shows the overall goodness of fit of the two PC-R models fit to the 1994 NAEP data extract. A likelihood-ratio test statistic indicates that the model goodness of fit to the data significantly improves when item-specific rater effects are taken into account ($\chi^2_{256} = 374.0$, $p < 0.01$). This result is consistent with our analyses of the 1992 extract described in 5.3 above. According to the AIC index the regular PCM seems to be the better fitting model. The difference between the two criteria for both data sets may be due to the relatively small size of the samples that we are using relative to the size of the entire NAEP data sets. Had large enough data sets been used, it is likely that, for this data set, the model with rater-by-item parameters would be deemed better fitting by both criteria. In any case, it is the size of the rater-by-item effects that will determine their significance.

Figure 13 graphically displays the item-specific rater effects for the 1994 data in the logit scale. Four clusters of items were distinguished in the data set because they were rated by different sets of raters. Severity estimates are shown for the raters that rated the items in each of the clusters. For example, rater 52 varies considerably in severity for the items in cluster 2, being the most lenient rater on item 6, less lenient on item 11, fairly close to the average on items 7, 8, 9, and 12, and the most severe rater on item 10. The variability of rater 52 can be contrasted with the consistency of rater 40, who rated four items (6, 7, 9, and 11) fairly close to the average.

To make interpretation of the rater effects easier and to indicate their impact on a subject's raw score, selected rater effects are transformed and plotted in the raw score metric in Figure 14. This figure indicates how much the score expected for an average ability student on a particular item when rated by a particular rater deviates from the score expected for an average ability student on average (i.e., rated by the average rater). Confidence intervals are indicated around this mean deviation. When the confidence interval does not include zero, the rater is either significantly more severe (bar below the zero line) or significantly more lenient (bar above the zero line). We can see,

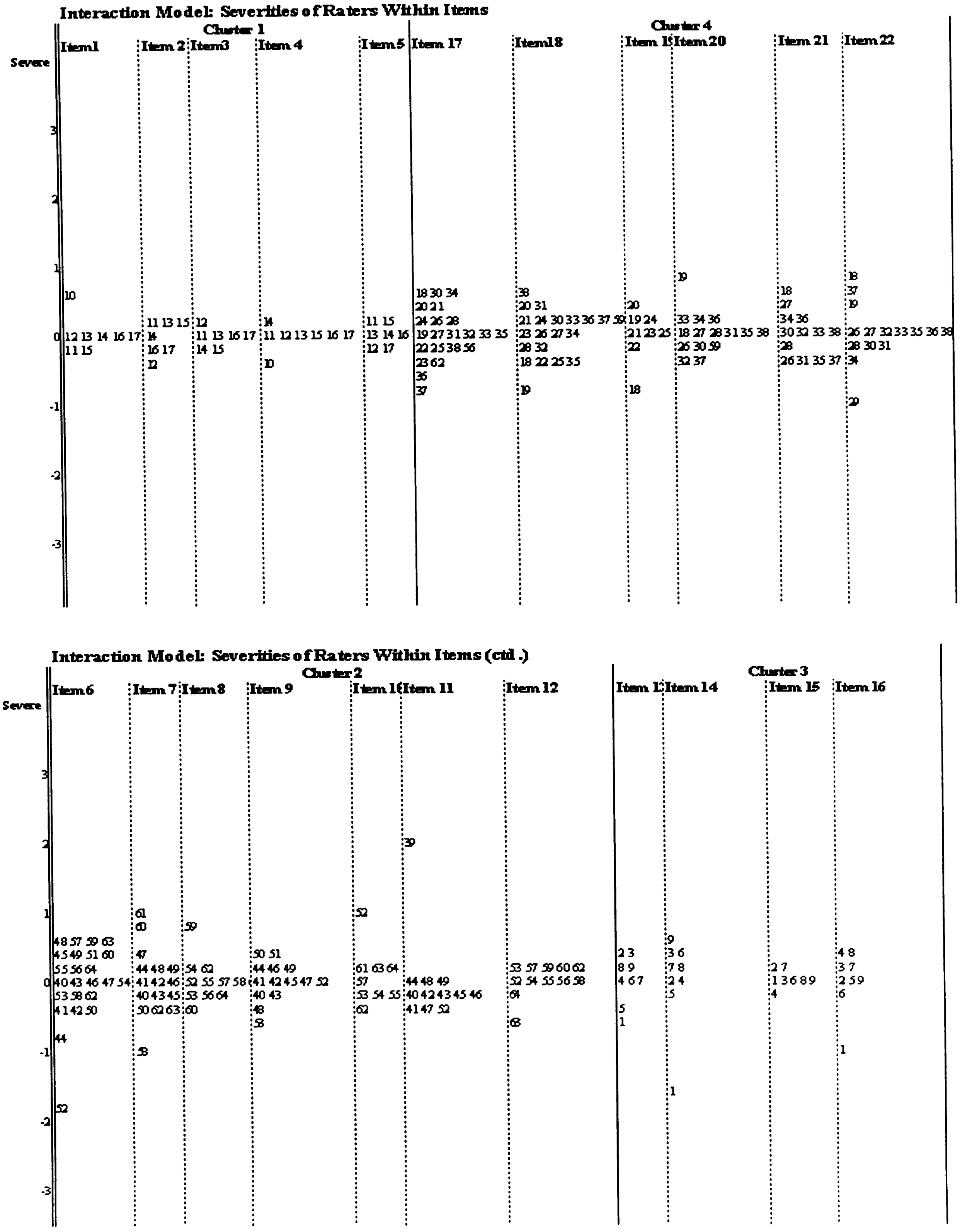


Figure 13: Distribution of estimated rater severities from the item-specific rater effect PC-R model in the 1994 NAEP State Assessment in Reading at grade 4.

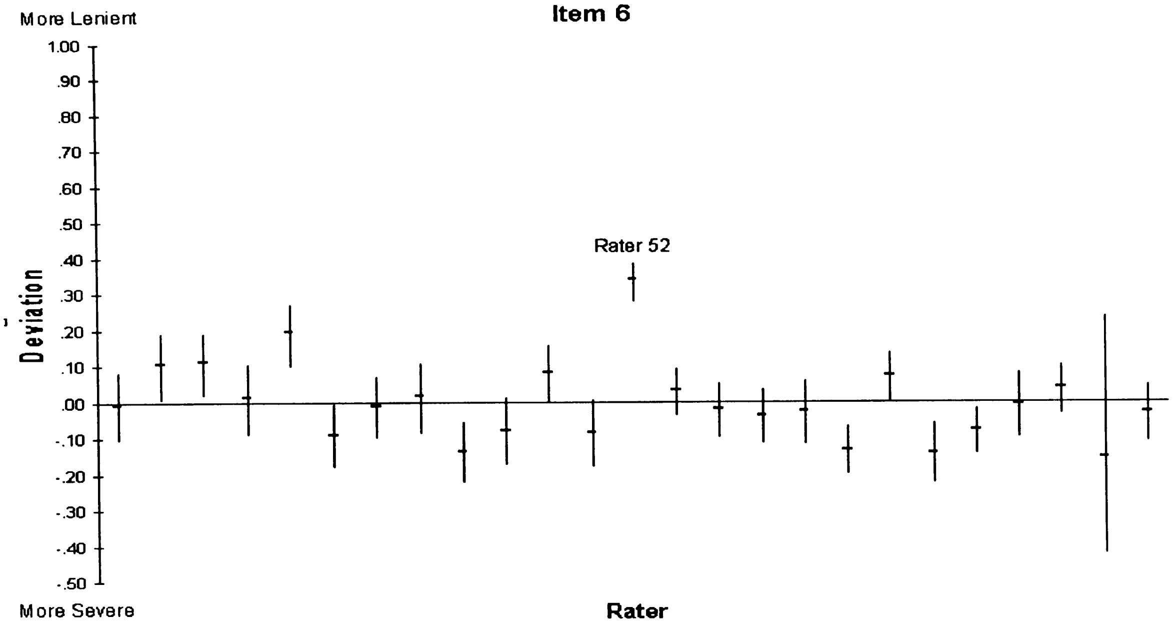


Figure 14: PC-R estimated score deviations for an average ability student when rated by each rater, as compared to the average rater, for item R012104 in NAEP's 1994 State Assessment Program in Reading at grade 4.

for example, how the leniency effect of rater 52 on item 6 translates into raw score differences for the persons the rater scored: roughly 30% of the subjects in a typical sample are more likely to get a score of one on this item when rated by rater 52 compared to a score of zero when rated by the average rater.

Overall, the estimated rater-by-item parameters have a standard deviation of 0.36, which is similar to the estimate of 0.33 found in the 1992 data set. However, the unbalanced nature of the distribution of item responses to raters (see Figure 12) makes interpretation of this number difficult. The question of the match of the PC-R model results to the GPC-R results arises here also. The ICCs in Figure 11 illustrate that the match is quite close.

As can be seen by comparing Figures 13 and 14 with Figures 1, 3, and 7 in earlier sections, the rater effects in NAEP are of a similar size to those observed elsewhere. Without more detailed data being made available, it is not possible to go beyond this. However, the similarity in size of the effects would lead one to speculate 1) that the impact on individual student results in NAEP would be similarly large, 2) that NAEP rater effects would also vary within scoring sessions, and 3) that NAEP rater effects may be reduced by feedback strategies.

7 Quantifying and minimizing rater effects by design

As we have noted above, classical indices of rater reliability alone are inadequate for describing the impact of rater errors when an IRT scale is used for scoring, as is the case in NAEP. In this section we investigate the impact of rater effects on IRT scale scores and classical test reliability. NAEP, of course, does not report individual test scale scores. Instead, five realizations from the posterior distribution for individual scale scores (i.e., plausible values) are generated and used to calculate NAEP statistics. Replicating NAEP's plausible value generation by fitting NAEP's conditioning model is beyond the scope of the present study. It is nonetheless instructive to investigate the relationship between IRT scale scores and rater effects like those that have been or might be observed in NAEP or other educational assessments containing constructed-response items. We present the results of such an investigation in this section.

We have also considered other strategies such as the rater feedback strategies described in section 4. Unfortunately, the data made available to us, and to researchers in general (i.e., the 1992 and 1994 NAEP data CD-ROMs), do not give enough information to carry out interesting research beyond what we describe here, which is primarily descriptive of the problems. More interesting and useful work along those lines will have to await an increase in understanding of the nature of the problem by those who carry out the NAEP scoring, and a readiness to share their information with the general research community.

As noted in section 5.2 above, the impact of rater biases on test scores depends on both the nature of the biases and on the allocation design for assigning item responses to raters. We investigated these relationships by simulating NAEP responses under several configurations of rater error types and rater allocation designs.

To clarify the question of interest here—how additive rater effects in the logit scale affect test scores on the IRT scale and classical test reliabilities—we focus on one complete set of items presented to a subset of examinees. In particular, we consider scores that would be assigned to students responding to items in one particular NAEP test booklet, containing two blocks of items on NAEP's Literary Experience Reading scale. The particular booklet number is R3, as defined in the NAEP Technical Report (Mazzeo, Allen, and Kline, 1995, p. 31). This booklet contains 10 multiple-choice items and 10 constructed-response items.

Experimental Conditions:

- **Rater severity type** was classified in one of three conditions depending on the overall variance of rater-by-item effects and on the proportion of that variance attributable to overall severity/leniency of individual raters across items. The first condition represents a control condition—no rater effects are present, and data are generated from a standard GPC model. The second rater effect condition approximately reproduces the nature of the rater effects observed in the earlier NAEP analyses (in sections 5 and 6). Under this "mild" condition the standard deviation, $\sigma_{\rho_{rj}}$, of the rater-by-item effects was 0.34, and 54% of the variance is attributable to overall (mean) rater effects across items (i.e., $\sigma^2_{\bar{\rho}_{r\cdot}} = 0.54\sigma^2_{\rho_{rj}}$). The third condition represents a very severe rater variability ($\sigma_{\rho_{rj}} = 1.43$) which is almost entirely accounted for by overall rater severity: $\sigma^2_{\bar{\rho}_{r\cdot}} = 0.98\sigma^2_{\rho_{rj}}$.

- **Allocation design** had three conditions. In 1992 NAEP raters were randomly assigned to student papers, but one rater scores all performances by the student. This assignment "by student" is the first allocation design condition. The second condition reflects the practice of NAEP in 1994. In this "random" condition, raters are randomly assigned to student responses, so each response by each student is rated by a randomly selected rater. The third allocation design is proposed as a way to systematically cancel out the effects of any rater bias at the test booklet level. In this "stratified" condition, the set of raters are divided into ten deciles based on rater severity, separately for each item. Each of the ten open-ended responses of a student are then distributed randomly so that one rater from each severity decile rates one response. This design eliminates the possibility that a booklet will by chance be rated by a preponderance of severe (or lenient) raters. It is important to note that in this simulation we assume that the rater severities are known (see the discussion below).

It is also important to clarify what is being simulated and what is being held fixed in this simulation study. The following values are held fixed: First, there are $N = 1,000$ examinees with proficiencies θ's fixed at 100 equally spaced quantiles of a $N(0,1)$ distribution. There are ten students at each unique θ, and the set of θ's are consistent with a $N(0,1)$ distribution. Second, there are $J = 20$ NAEP items with parameters as reported in the NAEP Technical Report (Mazzeo, Allen, and Kline, 1995, p. 323). Ten are multiple-choice items, four are two-level constructed-response items, four are 3-level constructed-response items, and two are 4-level constructed-response items. Third, there are $R = 20$ raters with severities ρ_{rj} fixed at equally spaced quantiles of the normal distribution implied by their experimental condition as described above. The number of NAEP raters scoring any single item ranged from 7 to 26, and this irregularity, as well as the highly uneven number of ratings made by any rater are problematic in reality (see section 6.2) and not replicated in our simulation.

The following values are simulated: First, the allocation of responses to raters is carried out randomly according to allocation design. Second, rated item responses are randomly generated based on the proficiency, item, and rater parameters. Third, maximum likelihood θ estimates are obtained for each vector of responses using the IRT program PARDUX (Burket, 1996). Two replications of these data generation and θ estimation steps are performed for each fixed student-item combination, over which only the rater assignment is varied. The two replicated response vectors yield two raw scores (total number of points), and the correlation of these two raw scores provides an estimate of the classical test reliability. Each estimated θ may be compared to the true θ used to generate the data, and the square root of the mean squared error (RMSE) provides an estimate of the IRT standard error of measurement.

Finally, the entire simulation was conducted ten times under each condition, because this allows us to report not only the mean statistics but also the standard error of the mean, which quantifies the uncertainty attributable to the simulation process (i.e., the Monte Carlo standard error). Thus the results reported in Tables 10 through 13 are based on simulated responses of 10,000 examinees. Since the allocation design is irrelevant when no rater effects are present, results for severity type "none" are collapsed across allocation design and 30,000 simulated examinees are used in the calculation of RMSE and reliability.

Of primary interest in the simulated data sets are the following:

- Accuracy of the resulting scale scores, as measured by the RMSE for estimated and true θ's.
- Classical test reliability, as measured by the correlation of the two replicated raw scores.

7.1 Simulation results

Table 10 presents estimates of classical reliability for each experimental condition. These estimates are means across 10 replications of each 1,000-examinee simulation described above. Standard errors associated with these means are given in parentheses. Two estimated reliabilities may be viewed as significantly different (i.e., well distinguished from each other by this estimation method) if the roughly 4-standard-error-wide intervals centered at the estimates do not overlap.

The reliability of the test booklet raw score is 0.863 when no rater effects are present, and this reliability drops significantly under "mild" rater effects in a "by student" design, and under "severe" rater effects in any of the three allocation designs. A significant decrease in reliability is avoided

Allocation	Severity Type		
	None	Mild	Severe
By student	0.863 (0.001)	0.854 (0.002)	0.717 (0.006)
Random		0.861 (0.002)	0.838 (0.002)
Stratified		0.864 (0.002)	0.849 (0.001)

Table 10: Estimated classical reliability coefficients from a set of responses to items in one 20-item NAEP test booklet, for several types of rater effects and rater allocation schemes. Standard errors of the estimates are in parentheses.

under "mild" rater effects if the allocation is either "random" or "stratified." The increase in reliability gained by randomization in the presence of "mild" rater effects from 0.854 to 0.861, may be thought of as equivalent to an increase in test booklet length of 6%, or about 1 "average" item (using the Spearman-Brown formula; see, e.g., Allen and Yen, 1979, p. 86). Since the "mild" rater effect is approximately that observed in NAEP, we estimate that by switching from a "by student" design in 1992 to a "random" design in 1994, NAEP gained a measure of accuracy approximately equivalent to an increase in test booklet length of one item.

The stratified randomization allocation design provides a significant increase in reliability in the presence of known, severe rater effects. We stress that these rater effects are quite severe, and that we are assuming them to be known. We consider simulations under "severe" rater effects and "stratified" allocation design to be proof of a promising concept. General usefulness of such an approach will depend on our ability to make accurate, real-time estimates of rater severity.

Table 11 presents the square root of the mean squared error (RMSE) in estimating θ based on simulated responses to the complete NAEP test booklet, for each level of rater severity and each allocation design. The pattern of differences is consistent with those observed among the reliabilities, with one notable anomaly: the RMSE for a stratified randomization under severe rater effects is actually lower than the RMSE attained when no rater effects are present. Further investigation reveals that the stratification results in smaller standard deviations for both realized raw scores and estimated scale scores (approximately 5% in each case), and consequently results in smaller RMSE without necessarily improved reliability. Viewed in this light, the smaller RMSE under stratification is similar to what one would expect from a shrinkage estimator, and thus RMSE should not be the considered as the sole basis for comparison of methodologies. We note again that classical reliability remains lower under "severe" rater effects even under stratified allocation.

We also repeated the complete simulation study described above using only the constructed-

	Severity Type		
Allocation	None	Mild	Severe
By student	0.480 (0.001)	0.495 (0.003)	0.775 (0.005)
Random		0.481 (0.003)	0.494 (0.003)
Stratified		0.479 (0.003)	0.463 (0.003)

Table 11: Square root of the mean squared error in estimating θ from the 20-item NAEP test booklet, for several types of rater effects and rater allocation schemes.

	Severity Type		
Allocation	None	Mild	Severe
By student	0.808 (0.001)	0.798 (0.002)	0.590 (0.006)
Random		0.804 (0.003)	0.770 (0.003)
Stratified		0.804 (0.003)	0.785 (0.003)

Table 12: Estimated classical reliability coefficients from the abbreviated test containing only the ten constructed-response items in the NAEP test booklet. In the absence of multiple-choice items, the impact of systematic rater effects is exacerbated.

response items from NAEP test booklet R3. The results for reliabilities and RMSE are presented in Tables 12 and 13. We can see by comparing Tables 12 and 10 that the benefits of randomization under severe rater effects are proportionately greater for tests consisting of only constructed-response items. It is under these conditions, too, that a stratified randomization brings the greatest improvement. Figure 15 compares estimated IRT standard error curves under regular randomization and stratified randomization. The improvement in reliability is estimated to be equivalent to a 9% increase in test length.

	Severity Type		
Allocation	None	Mild	Severe
By student	0.582 (0.002)	0.604 (0.004)	0.971 (0.007)
Random		0.586 (0.004)	0.607 (0.004)
Stratified		0.587 (0.003)	0.562 (0.004)

Table 13: RMSE in estimating θ using only the ten constructed-response items from the 20-item NAEP test booklet.

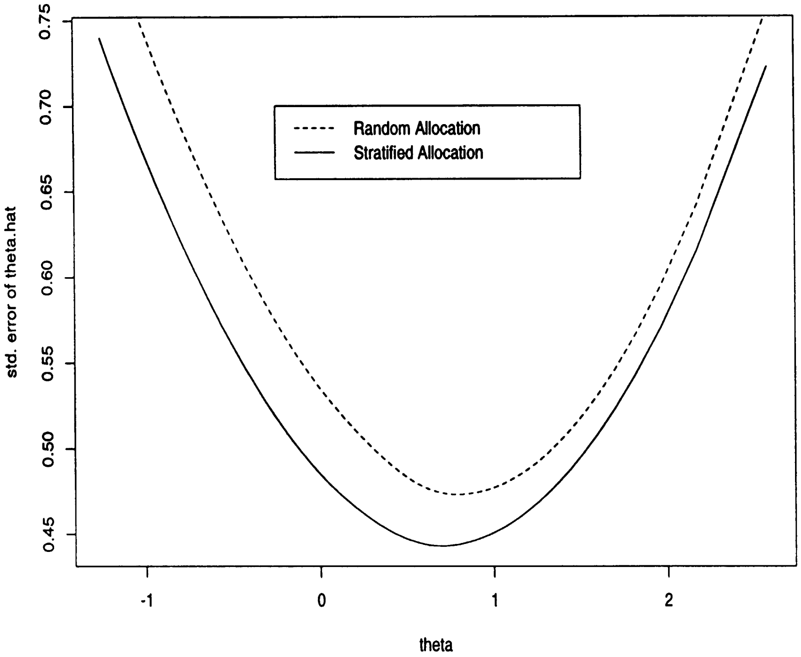


Figure 15: Estimated standard error of measurment (SEM) curves for the ten constructed- response items in booklet R3, in the presence of severe rater effects, under two allocation designs. Stratified randomization results in an improvement over simple randomization equivalent to a 9% increase in test length.

8 Conclusions and recommendations

8.1 Conclusions

The professional scoring for the NAEP open-ended items is of "industry standard" quality. This is clear by noting the similar range for the 1994 "matched pair" scores with, say, the CLAS Mathematics results noted above. The matches for the NAEP dichotomous items are somewhat higher than the CLAS matches, and the matches for the items with more score levels are the same or a little lower. We have characterized these discrepancies as rater biases (rater "severities"), and past research on CLAS and the Golden State Exam have been used to demonstrate that matches at the "industry standard" level may hide within them some large and troubling effects. When these results are aggregated, so long as the raters are well distributed across students within groups, this bias will usually be reduced, or even eliminated. However, the rater effects will persist, at least in theory, in the form of an underestimation of error variance. In the NAEP context, this will emerge as an underestimation in plausible value variance, which will affect secondary analyses, making any inferences based on the plausible values less conservative than they should be.

There have been suggestions of changes to NAEP that would affect this argument. For example, it has been suggested that NAEP should include a component that examines students' progress through the school years (Greeno, Pearson, and Schoenfeld, 1996). If this were to be a serious consideration, then the reduction in bias due to aggregation would not be relevant, and one would have to deal more directly with rater effects. This would, of course, be exacerbated if rater training and control characteristics varied from year to year, an effect we could not study with the current state of data recording in NAEP (see more on this below).

A review of the literature reveals that rater effects can be quite significant, and that they may take several forms. Rater bias is present when individual raters have consistent tendencies to be differentially severe or lenient in rating particular test items. Raters may also drift, becoming more harsh or lenient over the course of the rating period. The magnitude of rater effects and their impact on test scores can be quite significant, and yet this may be well hidden when only a few traditional measures of reliability (e.g., percent exact agreement among raters) are reported. That is, it is quite possible to have high percentages of exact agreement between raters and yet have significant amounts of rater bias affecting test scores.

Providing raters with periodic feedback during the rating process can significantly improve the quality of ratings, although effective intervention requires fast and accurate algorithms for quantifying rater severity.

Analyses of data from 1992 and 1994 NAEP State Reading Assessments at grade 4 reveal several important facts about rater effects in NAEP. Rater effects, in particular, differential severity of raters scoring individual items, are detectable in NAEP. Quantifying the size and impact of these effects is hampered by several factors, two of the most important being that 1) the technology for generalizing NAEP's scaling models to include rater parameters is currently in its formative stages, and 2) the design for the allocation of responses to raters is unbalanced. Our analyses address and partially overcome the first limitation; the second limitation can and should be addressed in the design of future NAEP scoring sessions.

The within-year rater effects we detect in NAEP are not particularly large, especially when considered in light of other sources of uncertainty and error in NAEP. In the context of NAEP, these rater effects are mitigated by 1) the presence of multiple-choice items in addition to constructed-response items, 2) the randomization of individual responses to raters, and 3) the aggregate nature of NAEP's reported statistics. In this context, the across-year rater effects may be of more importance.

The method of distributing responses to raters can have very significant consequences for the impact of rater errors. We found that randomization of individual responses instead of intact booklets may lead to a significant reduction in the error associated with estimated proficiencies. This improvement is especially significant in the presence of large rater biases that tend to be consistent across the items of a test. This item-by-item randomization, not used in 1992 NAEP but adopted for 1994 NAEP, leads to an improvement in the accuracy of plausible values that we estimate to be equivalent to adding one additional test item to NAEP's roughly 20-item test booklets.

We introduced a *stratified* randomization procedure that attempts to cancel the residual rater biases at a test score (or plausible values) level. This procedure, which could be incorporated into an integrated system for rater training, monitoring, and feedback, is shown in simulations to significantly improve proficiency estimation in the presence of severe rater effects. This finding is of general interest to the educational measurement field and should be investigated further and tested on a pilot basis. Implementation of such a strategy depends on the implementation of rater monitoring methods such as those described above.

The randomization of responses to raters needs to be carried out in a way that ensures that unbalanced designs do not result. Regardless of which particular randomization procedure is used, the distribution of responses to raters should be conducted in a statistically balanced fashion.

NAEP rescores 25% of the responses to open-ended items. Currently, information from the second ratings is used only for quality control purposes. Once levels of exact agreement between ratings are deemed acceptably high, the second rating is discarded and the first is retained and used for subsequent inference (see, e.g., Johnson, Mazzeo, and Kline, 1994, pp. 88–91). Information from the second set of ratings, if incorporated appropriately, should bring greater precision to NAEP's reported statistics. In generalizability theory, the inclusion of second ratings is a standard and accepted practice. The current methods for using second ratings in IRT have been criticized on the grounds that they overestimate the contribution of the repeated measures (Patz, 1996). The amount of additional information available to NAEP but not used should motivate useful development of appropriate statistical methodology for incorporating information from multiple ratings of student work.

8.2 Recommendations

Based on the analyses conducted in this project, a review of related literature, and experiences from related research projects on rater effects, we make the following recommendations for consideration by the National Assessment Governing Board in its redesign of NAEP:

1. NCES and NAEP should continue to develop a better framework for reporting on rater reliability in IRT contexts. In particular, NCES should require that NAEP contractors quantify how reported statistics would be expected to vary over replications of the professional scoring process.

2. NCES and its NAEP contractors should make more detailed information on the scoring process available, including time-stamped scoring data, read-behind, and/or check-sets data. This will facilitate investigation of the behavior of raters over the course of the scoring sessions and also from year to year.

3. NCES and its NAEP contractors should continue to develop and deploy systems that take full advantage of imaging technology in professional scoring. In particular, continued advances should be encouraged in systems for randomizing responses to raters, monitoring rater performance, and providing raters real-time feedback.

4. NCES should experiment with advanced randomization procedures based on real-time monitoring of rater severities in order to cancel residual differences in rater severities at the scale score (i.e., plausible values) level.

5. NCES should investigate improved methods of rubric standardization using imaging in order to increase the validity of NAEP's longitudinal equating.

6. NCES should encourage research to develop appropriate statistical methodology for incorporating information from multiple ratings of student work when item response theory scoring is used.

References

Allen, M. J., and Yen, W. M. (1979). Introduction to measurement theory. Monterey, CA: Brooks/Cole.

Burket, G. (1996). PARDUX [Computer software]. Monterey, CA: CTB/McGraw-Hill.

Cronbach, L. J., Linn, R. L., Brennan, R. L., and Haertel, E. (1995). Generalizability analysis for educational assessments. *Evaluation Comment.* Los Angeles: UCLA's Center for the Study of Evaluation and The National Center for Research on Evaluation, Standards and Student Testing. `http://www.cse.ucla.edu`.

CTB/McGraw-Hill. (1995). *Technical Report of the California Learning Assessment System, 1994.* Monterey, CA: Author.

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

Engelhard, G., Jr. (1996). Evaluating rater accuracy in performance assessments. *Journal of Educational Measurement, 33,* 56–70.

Fischer, G. H. (1973). The linear logistic model as an instrument in educational research. *Acta Psychologica, 37,* 359–374.

Fischer, G. H. (1983). Logistic latent trait models with linear constraints. *Psychometrika, 48,* 3–26.

Fischer, G. H., and Parzer, P. (1991). An extension of the rating scale model with an application to the measurement of change. *Psychometrika, 56,* 637–651.

Forsyth, R., Hambleton, R., Linn, R., Mislevy, R., and Yen, W. (1996). Design/Feasibility Team report to the National Assessment Governing Board.

Greeno, J.G., Pearson, P.D., and Schoenfeld, A.H. (1996). Implications for NAEP of research on learning and cognition. Research report, Institute for Research on Learning, Menlo Park, CA.

Hoskens, M., Wilson, M., and Stavisky, H. (1997). Accounting for rater effects in large scale testing using item response theory. Paper presented at the European meeting of the Psychometric Society, Spain.

Huguenard, B. R., Lerch, F. J., Junker, B. W., Patz, R. J., and Kass, R. E. (1997). Working memory failure in phone-based interaction. *ACM Transactions on Computer-Human Interaction, 4(2),* 67–102.

Johnson, E. G., Mazzeo, J., and Kline, D. L. (1994). *Technical Report of the NAEP 1992 Trial State Assessment Program in Reading.* Educational Testing Service and National Center for Education Statistics.

Jones, L. (1996). A history of the National Assessment of Educational Progress and some questions about its future. *Educational Researcher, 25*, 15–21.

Koretz, D., Stecher, B., Klein, S., and McCaffrey, D. (1994). The Vermont Portfolio Assessment Program: findings and implications. *Educational Measurement: Issues and Practice, 13,* 5–16.

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

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

Mazzeo, J., Allen, N. L., and Kline, D. L. (1995). *Technical Report of the NAEP 1994 Trial State Assessment Program in Reading.* Educational Testing Service and National Center for Education Statistics.

Muraki, E. (1992). A generalized partial credit model: Application of an EM algorithm. *Applied Psychological Measurement, 16*, 159–176.

Myford C. M., and Mislevy, R. J. (1995). *Monitoring and improving a portfolio assessment system.* Center for Performance Assessment Research Report. Princeton, NJ: Educational Testing Service.

Patz, R. J. (1996). Markov chain Monte Carlo methods for item response theory models with applications for the National Assessment of Educational Progress. Doctoral dissertation, Carnegie Mellon University.

Patz, R. J., and Junker, B. W. (1997). A straightforward approach to Markov chain Monte Carlo methods for item response models. Manuscript submitted for publication.

Patz, R. J., and Junker, B. W. (1997b). Applications and extensions of MCMC in IRT: multiple item types, missing data, and rated responses. Manuscript.

Ponocny, I., and Ponocny-Seliger, E. (in press). Applications of the program LpcM in the field of measuring change. In M. Wilson, G. Engelhard, and K. Draney, *Objective Measurement IV: Theory into Practice.* Norwood, NJ: Ablex.

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests.* Copenhagen: Danish Institute for Educational Research.

Wainer, H. (1993). Measurement problems. *Journal of Educational Measurement, 30,* 1–21.

Wilson, M., and Case, H. (1996, June). An investigation of the feasibility and potential effects of rater feedback on rater errors. Paper presented at the CCSSO Conference, Phoenix, AZ.

Wilson, M., and Wang, W. (1995). Complex composites: Issues that arise in combining different modes of assessment. *Applied Psychological Measurement, 19(1)*, 51–72.

Wu, M., Adams, R.J., and Wilson, M. (in press). ConQuest [Computer software]. Hawthorn, Australia: ACER.

Listing of NCES Working Papers to Date

Please contact Ruth R. Harris at (202) 219-1831
if you are interested in any of the following papers

Number	Title	Contact
94-01 (July)	Schools and Staffing Survey (SASS) Papers Presented at Meetings of the American Statistical Association	Dan Kasprzyk
94-02 (July)	Generalized Variance Estimate for Schools and Staffing Survey (SASS)	Dan Kasprzyk
94-03 (July)	1991 Schools and Staffing Survey (SASS) Reinterview Response Variance Report	Dan Kasprzyk
94-04 (July)	The Accuracy of Teachers' Self-reports on their Postsecondary Education: Teacher Transcript Study, Schools and Staffing Survey	Dan Kasprzyk
94-05 (July)	Cost-of-Education Differentials Across the States	William Fowler
94-06 (July)	Six Papers on Teachers from the 1990-91 Schools and Staffing Survey and Other Related Surveys	Dan Kasprzyk
94-07 (Nov.)	Data Comparability and Public Policy: New Interest in Public Library Data Papers Presented at Meetings of the American Statistical Association	Carrol Kindel
95-01 (Jan.)	Schools and Staffing Survey: 1994 Papers Presented at the 1994 Meeting of the American Statistical Association	Dan Kasprzyk
95-02 (Jan.)	QED Estimates of the 1990-91 Schools and Staffing Survey: Deriving and Comparing QED School Estimates with CCD Estimates	Dan Kasprzyk
95-03 (Jan.)	Schools and Staffing Survey: 1990-91 SASS Cross-Questionnaire Analysis	Dan Kasprzyk
95-04 (Jan.)	National Education Longitudinal Study of 1988: Second Follow-up Questionnaire Content Areas and Research Issues	Jeffrey Owings
95-05 (Jan.)	National Education Longitudinal Study of 1988: Conducting Trend Analyses of NLS-72, HS&B, and NELS:88 Seniors	Jeffrey Owings

Listing of NCES Working Papers to Date--Continued

Number	Title	Contact
95-06 (Jan.)	National Education Longitudinal Study of 1988: Conducting Cross-Cohort Comparisons Using HS&B, NAEP, and NELS:88 Academic Transcript Data	Jeffrey Owings
95-07 (Jan.)	National Education Longitudinal Study of 1988: Conducting Trend Analyses HS&B and NELS:88 Sophomore Cohort Dropouts	Jeffrey Owings
95-08 (Feb.)	CCD Adjustment to the 1990-91 SASS: A Comparison of Estimates	Dan Kasprzyk
95-09 (Feb.)	The Results of the 1993 Teacher List Validation Study (TLVS)	Dan Kasprzyk
95-10 (Feb.)	The Results of the 1991-92 Teacher Follow-up Survey (TFS) Reinterview and Extensive Reconciliation	Dan Kasprzyk
95-11 (Mar.)	Measuring Instruction, Curriculum Content, and Instructional Resources: The Status of Recent Work	Sharon Bobbitt & John Ralph
95-12 (Mar.)	Rural Education Data User's Guide	Samuel Peng
95-13 (Mar.)	Assessing Students with Disabilities and Limited English Proficiency	James Houser
95-14 (Mar.)	Empirical Evaluation of Social, Psychological, & Educational Construct Variables Used in NCES Surveys	Samuel Peng
95-15 (Apr.)	Classroom Instructional Processes: A Review of Existing Measurement Approaches and Their Applicability for the Teacher Follow-up Survey	Sharon Bobbitt
95-16 (Apr.)	Intersurvey Consistency in NCES Private School Surveys	Steven Kaufman
95-17 (May)	Estimates of Expenditures for Private K-12 Schools	Stephen Broughman
95-18 (Nov.)	An Agenda for Research on Teachers and Schools: Revisiting NCES' Schools and Staffing Survey	Dan Kasprzyk
96-01 (Jan.)	Methodological Issues in the Study of Teachers' Careers: Critical Features of a Truly Longitudinal Study	Dan Kasprzyk

Listing of NCES Working Papers to Date--Continued

Number	Title	Contact
96-02 (Feb.)	Schools and Staffing Survey (SASS): 1995 Selected papers presented at the 1995 Meeting of the American Statistical Association	Dan Kasprzyk
96-03 (Feb.)	National Education Longitudinal Study of 1988 (NELS:88) Research Framework and Issues	Jeffrey Owings
96-04 (Feb.)	Census Mapping Project/School District Data Book	Tai Phan
96-05 (Feb.)	Cognitive Research on the Teacher Listing Form for the Schools and Staffing Survey	Dan Kasprzyk
96-06 (Mar.)	The Schools and Staffing Survey (SASS) for 1998-99: Design Recommendations to Inform Broad Education Policy	Dan Kasprzyk
96-07 (Mar.)	Should SASS Measure Instructional Processes and Teacher Effectiveness?	Dan Kasprzyk
96-08 (Apr.)	How Accurate are Teacher Judgments of Students' Academic Performance?	Jerry West
96-09 (Apr.)	Making Data Relevant for Policy Discussions: Redesigning the School Administrator Questionnaire for the 1998-99 SASS	Dan Kasprzyk
96-10 (Apr.)	1998-99 Schools and Staffing Survey: Issues Related to Survey Depth	Dan Kasprzyk
96-11 (June)	Towards an Organizational Database on America's Schools: A Proposal for the Future of SASS, with comments on School Reform, Governance, and Finance	Dan Kasprzyk
96-12 (June)	Predictors of Retention, Transfer, and Attrition of Special and General Education Teachers: Data from the 1989 Teacher Followup Survey	Dan Kasprzyk
96-13 (June)	Estimation of Response Bias in the NHES:95 Adult Education Survey	Steven Kaufman
96-14 (June)	The 1995 National Household Education Survey: Reinterview Results for the Adult Education Component	Steven Kaufman

Listing of NCES Working Papers to Date--Continued

Number	Title	Contact
96-15 (June)	Nested Structures: District-Level Data in the Schools and Staffing Survey	Dan Kasprzyk
96-16 (June)	Strategies for Collecting Finance Data from Private Schools	Stephen Broughman
96-17 (July)	National Postsecondary Student Aid Study: 1996 Field Test Methodology Report	Andrew G. Malizio
96-18 (Aug.)	Assessment of Social Competence, Adaptive Behaviors, and Approaches to Learning with Young Children	Jerry West
96-19 (Oct.)	Assessment and Analysis of School-Level Expenditures	William Fowler
96-20 (Oct.)	1991 National Household Education Survey (NHES:91) Questionnaires: Screener, Early Childhood Education, and Adult Education	Kathryn Chandler
96-21 (Oct.)	1993 National Household Education Survey (NHES:93) Questionnaires: Screener, School Readiness, and School Safety and Discipline	Kathryn Chandler
96-22 (Oct.)	1995 National Household Education Survey (NHES:95) Questionnaires: Screener, Early Childhood Program Participation, and Adult Education	Kathryn Chandler
96-23 (Oct.)	Linking Student Data to SASS: Why, When, How	Dan Kasprzyk
96-24 (Oct.)	National Assessments of Teacher Quality	Dan Kasprzyk
96-25 (Oct.)	Measures of Inservice Professional Development: Suggested Items for the 1998-1999 Schools and Staffing Survey	Dan Kasprzyk
96-26 (Nov.)	Improving the Coverage of Private Elementary-Secondary Schools	Steven Kaufman
96-27 (Nov.)	Intersurvey Consistency in NCES Private School Surveys for 1993-94	Steven Kaufman

Listing of NCES Working Papers to Date--Continued

Number	Title	Contact
96-28 (Nov.)	Student Learning, Teaching Quality, and Professional Development: Theoretical Linkages, Current Measurement, and Recommendations for Future Data Collection	Mary Rollefson
96-29 (Nov.)	Undercoverage Bias in Estimates of Characteristics of Adults and 0- to 2-Year-Olds in the 1995 National Household Education Survey (NHES:95)	Kathryn Chandler
96-30 (Dec.)	Comparison of Estimates from the 1995 National Household Education Survey (NHES:95)	Kathryn Chandler
97-01 (Feb.)	Selected Papers on Education Surveys: Papers Presented at the 1996 Meeting of the American Statistical Association	Dan Kasprzyk
97-02 (Feb.)	Telephone Coverage Bias and Recorded Interviews in the 1993 National Household Education Survey (NHES:93)	Kathryn Chandler
97-03 (Feb.)	1991 and 1995 National Household Education Survey Questionnaires: NHES:91 Screener, NHES:91 Adult Education, NHES:95 Basic Screener, and NHES:95 Adult Education	Kathryn Chandler
97-04 (Feb.)	Design, Data Collection, Monitoring, Interview Administration Time, and Data Editing in the 1993 National Household Education Survey (NHES:93)	Kathryn Chandler
97-05 (Feb.)	Unit and Item Response, Weighting, and Imputation Procedures in the 1993 National Household Education Survey (NHES:93)	Kathryn Chandler
97-06 (Feb.)	Unit and Item Response, Weighting, and Imputation Procedures in the 1995 National Household Education Survey (NHES:95)	Kathryn Chandler
97-07 (Mar.)	The Determinants of Per-Pupil Expenditures in Private Elementary and Secondary Schools: An Exploratory Analysis	Stephen Broughman
97-08 (Mar.)	Design, Data Collection, Interview Timing, and Data Editing in the 1995 National Household Education Survey	Kathryn Chandler

Listing of NCES Working Papers to Date--Continued

Number	Title	Contact
97-09 (Apr.)	Status of Data on Crime and Violence in Schools: Final Report	Lee Hoffman
97-10 (Apr.)	Report of Cognitive Research on the Public and Private School Teacher Questionnaires for the Schools and Staffing Survey 1993-94 School Year	Dan Kasprzyk
97-11 (Apr.)	International Comparisons of Inservice Professional Development	Dan Kasprzyk
97-12 (Apr.)	Measuring School Reform: Recommendations for Future SASS Data Collection	Mary Rollefson
97-13 (Apr.)	Improving Data Quality in NCES: Database-to-Report Process	Susan Ahmed
97-14 (Apr.)	Optimal Choice of Periodicities for the Schools and Staffing Survey: Modeling and Analysis	Steven Kaufman
97-15 (May)	Customer Service Survey: Common Core of Data Coordinators	Lee Hoffman
97-16 (May)	International Education Expenditure Comparability Study: Final Report, Volume I	Shelley Burns
97-17 (May)	International Education Expenditure Comparability Study: Final Report, Volume II, Quantitative Analysis of Expenditure Comparability	Shelley Burns
97-18 (June)	Improving the Mail Return Rates of SASS Surveys: A Review of the Literature	Steven Kaufman
97-19 (June)	National Household Education Survey of 1995: Adult Education Course Coding Manual	Peter Stowe
97-20 (June)	National Household Education Survey of 1995: Adult Education Course Code Merge Files User's Guide	Peter Stowe
97-21 (June)	Statistics for Policymakers or Everything You Wanted to Know About Statistics But Thought You Could Never Understand	Susan Ahmed
97-22 (July)	Collection of Private School Finance Data: Development of a Questionnaire	Stephen Broughman

Listing of NCES Working Papers to Date--Continued

Number	Title	Contact
97-23 (July)	Further Cognitive Research on the Schools and Staffing Survey (SASS) Teacher Listing Form	Dan Kasprzyk
97-24 (Aug.)	Formulating a Design for the ECLS: A Review of Longitudinal Studies	Jerry West
97-25 (Aug.)	1996 National Household Education Survey (NHES:96) Questionnaires: Screener/Household and Library, Parent and Family Involvement in Education and Civic Involvement, Youth Civic Involvement, and Adult Civic Involvement	Kathryn Chandler
97-26 (Oct.)	Strategies for Improving Accuracy of Postsecondary Faculty Lists	Linda Zimbler
97-27 (Oct.)	Pilot Test of IPEDS Finance Survey	Peter Stowe
97-28 (Oct.)	Comparison of Estimates in the 1996 National Household Education Survey	Kathryn Chandler
97-29 (Oct.)	Can State Assessment Data be Used to Reduce State NAEP Sample Sizes?	Steven Gorman
97-30 (Oct.)	ACT's NAEP Redesign Project: Assessment Design is the Key to Useful and Stable Assessment Results	Steven Gorman
97-31 (Oct.)	NAEP Reconfigured: An Integrated Redesign of the National Assessment of Educational Progress	Steven Gorman
97-32 (Oct.)	Innovative Solutions to Intractable Large Scale Assessment (Problem 2: Background Questionnaires)	Steven Gorman
97-33 (Oct.)	Adult Literacy: An International Perspective	Marilyn Binkley
97-34 (Oct.)	Comparison of Estimates from the 1993 National Household Education Survey	Kathryn Chandler
97-35 (Oct.)	Design, Data Collection, Interview Administration Time, and Data Editing in the 1996 National Household Education Survey	Kathryn Chandler
97-36 (Oct.)	Measuring the Quality of Program Environments in Head Start and Other Early Childhood Programs: A Review and Recommendations for Future Research	Jerry West

Listing of NCES Working Papers to Date--Continued

Number	Title	Contact
97-37 (Nov.)	Optimal Rating Procedures and Methodology for NAEP Open-ended Items	Steven Gorman