DOCUMENT RESUME

ED 385 544 TM 023 960

AUTHOR Lukhele, Robert; And Others
TITLE On the Relative Value of Multiple-Choice, Constructed Response, and Examinee-Selected Items on Two Achievement Tests. Program Statistics Research Technical Report No. 93-28.
INSTITUTION Educational Testing Service, Princeton, N.J.
REPORT NO ETS-RR-93-6
PUB DATE Feb 93
NOTE 28p.
PUB TYPE Reports - Research/Technical (143)

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS *Achievement Tests; Advanced Placement; *Chemistry; *Constructed Response; High Schools; *High School Students; Item Response Theory; *Multiple Choice Tests; Scoring; Selection; Test Items; Test Use; *United States History
IDENTIFIERS Advanced Placement Examinations (CEEB)

ABSTRACT

Analyses based on fitting item response models to data from the College Board's Advanced Placement exams in Chemistry and United States History indicated that the constructed-response portion of the tests yielded little information over and above that provided by the multiple-choice sections. These tests also allow examinees to select subsets of the constructed-response items. Data from the operational administration of the 1989 Advanced Placement Test in Chemistry (taken by 18,462 students) and the 1988 Advanced Placement Test in United States History (taken by 82,842 students) were analyzed. It was found that scoring on the basis of the selections themselves provided almost as much information as did scoring on the basis of the answers. It was also determined that the chemistry test was too difficult for its primary goal, but that this could be at least partially corrected by taking into account information in the wrong answers to the multiple-choice items. Seven figures and two tables illustrate the discussion. (Contains 32 references.) (Author/SLD)

RR-93-6

On the Relative Value of Multiple-Choice, Constructed Response, and Examinee-Selected Items on Two Achievement Tests

Robert Lukhele
University of California

David Thissen
University of North Carolina

Howard Wainer
Educational Testing Service

PROGRAM STATISTICS RESEARCH

TECHNICAL REPORT NO. 93-28

Educational Testing Service
Princeton, New Jersey 08541

The Program Statistics Research Technical Report Series is designed to make the working papers of the Research Statistics Group at Educational Testing Service generally available. The series consists of reports by the members of the Research Statistics Group as well as their external and visiting statistical consultants.

On the Relative Value of Multiple-Choice, Constructed Response, and Examinee-Selected Items on Two Achievement Tests

Robert Lukhele
University of California

David Thissen
University of North Carolina

Howard Wainer
Educational Testing Service

Program Statistics Research
Technical Report No. 93-28

Research Report No. 93-6

Educational Testing Service
Princeton, New Jersey 08541

February 1993

On the relative value of multiple-choice, constructed-response, and examinee-selected items on two achievement tests[1]

Robert Lukhele
University of California
Santa Barbara

David Thissen
University of North Carolina
Chapel Hill

Howard Wainer
Educational Testing Service
Princeton

Abstract

Using analyses based on fitting item response models to data from the College Board's Advanced Placement exams in Chemistry and United States History, we found that the constructed-response portion of the tests yielded little information over and above that provided by the multiple-choice sections. These tests also allow examinees to select subsets of the constructed-response items; we found that scoring on the basis of the selections themselves provided almost as much information as did scoring on the basis of the answers. Finally, we found that the chemistry test was too difficult for its primary goal, but that this could be at least partially corrected by taking into account information in the wrong answers to the multiple-choice items.

---

[1]This work was collaborative in every respect and the order of the authors is alphabetical. The first author's work was performed while he was a Sumner Predoctoral Fellow at the Educational Testing Service. Thissen and Wainer's efforts were partially supported by the GRE Research Board; we are pleased to have this opportunity to acknowledge it. We are grateful to Jim Ferris, Nick Longford, Robert Mislevy, Rick Morgan, and Neal Thomas for helpful conversations, although they should not be held responsible for what we have done with their good advice. We would like to express special thanks to Jim Ramsay, whose test analysis program TESTGRAF provided us with the impetus to do the analyses that eventually resulted in Figure 7.

I. Introduction

One goal of the Advanced Placement (AP) Testing program of the College Board is to determine the proficiency attained by high school students in college level courses. Such courses are offered at more than 40% of the twenty-one thousand secondary schools in this country, including 80% of those having a junior class larger than 250 students. With few exceptions, AP exams are made up of one section consisting of multiple-choice (MC) items, and a second section of constructed-response (CR) items that are scored by carefully trained expert raters. The two section scores are then combined into a single score and a judgment is made, based solely on this score, about whether to grant college credit to the examinee.

Constructed-response items are expensive. They typically require a great deal of time from the examinee to answer and they cost a lot to score. In the AP testing program it was found (Wainer & Thissen, 1993) that a constructed-response test of equivalent reliability to a multiple-choice test takes from 4 to 40 times as long to administer and is typically hundreds to thousands of times more expensive to score. Given these practical drawbacks, why are constructed-response items included in tests? One reason is because they are thought to measure something different than multiple-choice tests, and what they measure is important. Before examining the validity of this claim, let us briefly describe some of the arguments supporting the use of constructed-response items.

The term constructed-response is generally used to refer to any question format that requires the test taker to produce a response in any way other than selecting from a list of alternative answers. Constructed-response tasks may be as simple as a short-answer question, or adding an arrow to a diagram. They may require the test taker to organize and write an essay, solve a multistep mathematics problem, draw a diagram or graph, or write an explanation of a procedure. The category even covers formats for evaluating musical performance, visual arts portfolios, and fluency in foreign languages (Pollack, Rock, & Jenkins, 1992).

Recent developments in cognitive theory suggest that new achievement tests must measure four important dimensions of performance (Glaser, 1985):

1. The extent to which an examinee's performance on the test is principled, that is, derived from interconnected rules rather than fragmentary pieces of information.

2. The size and direction of dynamic changes in students' strategies, which are hypothesized to reflect the structure of the mental models measured in (1).

3. The structure or representation of knowledge and cognitive skills.

4. The amount of automaticity of performance skills.

Multiple-choice items are suitable for measuring static knowledge (Tatsuoka, 1991), but have been challenged as being inadequate to fully assess these dimensions of cognitive performance. The limited opportunity for demonstrating in-depth knowledge afforded by this format, as well as the possibility that "test-wiseness" can contaminate the measurement, have prompted a search for alternatives to multiple-choice testing (Pollack, Rock, & Jenkins, 1992). Constructed-response items are thought to offer such an alternative.

The primary motivation for the use of constructed-response formats thus stems from the idea that they can measure traits that cannot be tapped by multiple-choice items, for example, dynamic cognitive processes (Bennett, Ward, Rock, & Lahart, 1990), educational assessment (Fiske, 1990; Fredericksen & Collins, 1989; Guthrie, 1984; Nickerson, 1989), identification of students' misconceptions in diagnostic testing (Birenbaum & Tatsuoka, 1987) and communicating to teachers and students the importance of practicing these "real-world" tasks (Sebrechts, Bennett & Rock, 1991).

Constructed-response questions are thought to replicate more faithfully the tasks examinees face in academic and work settings. Thus, there is an indirect benefit of constructed-response test formats that has nothing to do with their measurement characteristics. Because large-scale tests are highly visible within the education community, their content may provide cues to teachers about what is important to teach, and to students about what is important to learn. If it is known that students will be required to demonstrate competence in problem solving, graphing, verbal expression, essay organization and writing, these skills may be more likely to be emphasized in the classroom (Pollack, Rock, & Jenkins, 1992).

These arguments are powerful and convincing. It is our purpose, in this investigation, to measure the extent to which the constructed-response format, as represented in the AP testing program, provides the benefits expected of it. We know that it incurs the costs.

Why confine ourselves to the College Board's Advanced Placement Program? There are many reasons. Principal among them is because it is a very large program with hundreds of thousands of students participating in it annually. Students' performance on these tests has serious consequences; a single college course can cost $1,500 which can be used for other purposes if college credit is granted. Because of the serious and extensive character of these tests, considerable effort and great care is devoted to making them fair and valid[2]. We believe that the AP tests are sufficiently similar to other high quality tests that combine MC and CR items, to allow reasonably broad generalization of

---

[2]To say that "no expense is spared" in building these tests may be a bit hyperbolic, but not much. We know of several low volume tests (i.e. AP Music) that annually cost several times more to administer and score then the revenue they generate. Such altruistic concern for quality regardless of cost should not be denigrated.

our findings to other situations. In this paper we explore the value of some of the assumptions, both explicit and implicit, that underlie the structure of such tests.

We confine the rest of our discussion to the AP Chemistry and the AP United States History tests, since their quality is representative of many high quality tests. Each of these two tests manifests two characteristics of special interest to modern test development. These are:

- They contain both multiple-choice and constructed-response items
- They allow examinees a choice of which constructed-response items they will answer.

These two characteristics raise a variety of difficult psychometric and theoretical problems. Four of these are:

1. Does it make sense to combine the two different kinds of items into a single score? How can we tell? (Thissen, Wainer & Wang, 1993)
2. If we can combine them, how should we weight them? (Wainer & Thissen, 1993)
3. What is the relative value of these two different kinds of items? Under what circumstances should one be preferred to the other? How do they complement each other?
4. How do we score a test that allows choice? What assumptions are necessary? How likely are they to be true? How sensitive are our results to the accuracy of these assumptions? (Wainer, Wang & Thissen, 1991; Wang, Thissen & Wainer, 1993)

The references next to questions 1, 2, and 4 contain detailed examinations of the associated question, although each of these studies touches to a varying degree on each of the other questions. While this paper focuses on the third question, it too touches on aspects of the others.

The two tests under consideration differ in the way that the constructed-response portion of the test is scored. The CR items on the Chemistry test are analytically scored, whereas on the History test they are holistically scored. Thus, in some sense the constructed-response portions of these two tests represent the extremes on a continuum of scoring rubrics.

II. Data and Procedures

In this study we analyze all of the data from the operational administration of the 1989 Advanced Placement Test in Chemistry as well as from the 1988 administration of the Advanced Placement Test in United States History. For reasons of parsimony, we

shall report the analyses and results from the Chemistry test in detail and then summarize the results of parallel analyses of the History test.

The chemistry test is divided into two sections with 90 minutes allotted for each. Section I consists of 75 five-option multiple-choice questions and accounts for 45% of the total grade. Section II consists of problems and essay questions, and has four parts:

- Part A is a single problem (Problem 1) that all examinees must answer, and accounts for 14% of the total grade.
- Part B has two problems (Problems 2 and 3), and the examinee must answer exactly one of those. This part accounts for 14% of the total grade.
- Part C is treated as a single problem (Problem 4), but has eight components, from which the examinee must answer five. This part accounts for 8% of the total grade.
- Part D has five problems (Problems 5, 6, 7, 8 and 9) of which the examinee must answer three. This part accounts for 19% of the total grade.

Figure 1 shows the structure of the test graphically[3].

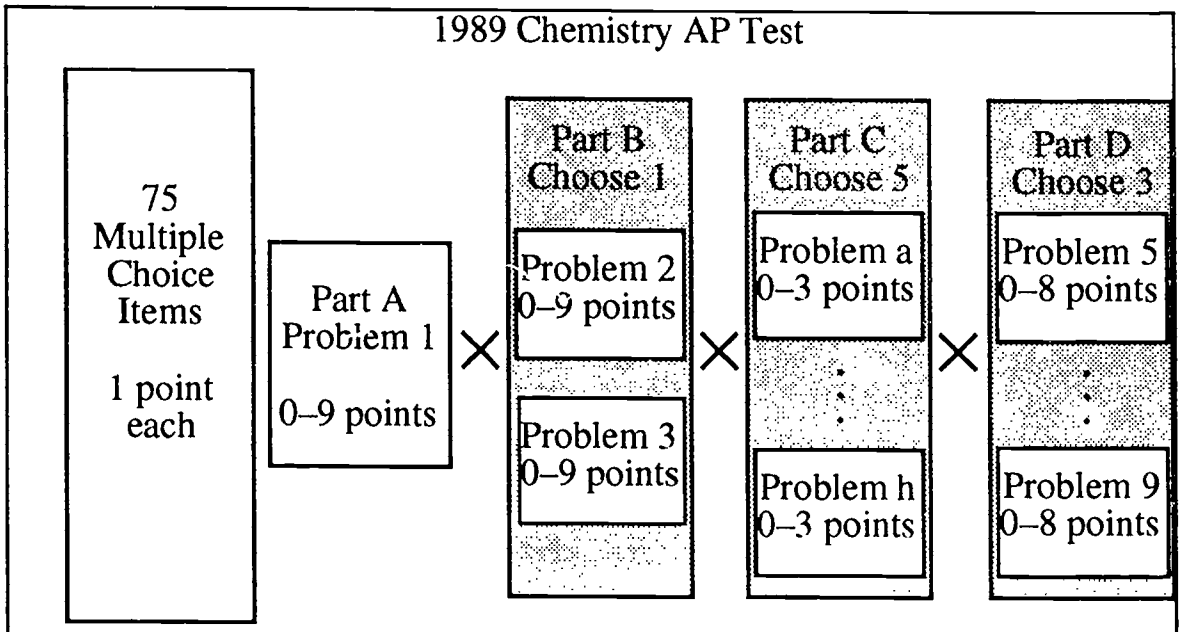

Figure 1. Graphical representation of the constituent parts of the College Board's Advanced Placement Test in Chemistry in 1989

[3]The "points" shown on Figure 1 relates to how each of the items is scored. Weights are then assigned to each part to yield the total score. The combination of points and weights yields the relative combinations reported above.

This form of the exam was taken by 18,462 students in 1989. The test form has been released and interested readers may obtain copies with the answers and a full description of the scoring methods from the College Board.

Parts B, C and D all contain choice. However, the choices in Part C were not recorded into the data base; the only data available were the scores obtained from answering five of these questions. A maximum of three points was given to each of these five questions. Thus, we treated Part C as a single question scored on a 16 point scale (0-15) without choice.

We constructed four versions of the AP Chemistry test. These tests were:

Test 75 — This test consisted of the 75 multiple-choice items.

Test 77 — Test 75 plus Problems 1 and 4.

Test 79 — Test 77 plus two nominal items; a dichotomous item indicating which problem in Part B was chosen and a polytomous item indicating which three problems in Part D were chosen[4].

Test 84 — Test 77 plus the seven choice items contained in Parts B and D. Of course only 5 of these seven items were ever answered by any examinee, yet all were available.

All of the data to score 75, 77 and 79 were observable. In the scoring of form 84 we assumed initially that the responses that were missing would have conformed to the same item tracelines as those that were observed. In the terminology made popular by Little & Rubin (1987), we assumed that the nonresponse was ignorable, conditioned on proficiency.

The measurement models

For each test, an appropriate model was fitted to the observed data using the maximum marginal likelihood algorithm described by Bock & Aitken (1981), as implemented in the computer program Multilog (Thissen, 1991; see also Thissen, 1982 and Thissen & Steinberg, 1984). That is, for each test, the item parameter estimates maximize the likelihood

---

[4]This polytomous item was treated as a nominal item with ten categories associated with the (5 choose 3 = 10) possible choices.

$$L = \prod_{response\ patterns} P(\mathbf{x})^{r_{\mathbf{x}}} \quad ,$$

in which $r_{\mathbf{x}}$ is the frequency of response pattern $\mathbf{x}$ and $P(\mathbf{x})$ represents the probability of that response pattern in the data,

$$P(\mathbf{x}) = \int \prod_{j=1}^{\#\ of\ items} T_j(x_j|\theta)\ \phi(\theta)\ d\theta \quad .$$

The latent variable (proficiency) is denoted θ, the trace lines describing the probability of a response in category x_j for item j as a function of θ, $T_j(x_j|\theta)$, take different functional forms for the different item types and response formats, and the population distribution for θ, $\phi(\theta)$, is assumed to be $N(0,1)$.

For the multiple-choice items, $x = 1$ (correct) or $x = 0$ (incorrect), and we use the conventional three-parameter logistic model (Lord, 1980), in which

$$T(x = 1) = c + \left[\frac{1 - c}{1 + \exp[-1.7a(\theta - b)]}\right] \quad .$$

There are three parameters for each item: a, reflecting the item's discrimination, b, reflecting the item's difficulty, and c, which is the probability that a low-proficiency test-taker responds correctly ("guessing").

For the constructed-response items, we use Samejima's (1969) graded model for ordered responses $x = k$, $k = 1, 2, \ldots, m$, where m reflects highest score:

$$T(x = k) = \frac{1}{1 + \exp[-a(\theta - b_{k-1})]} - \frac{1}{1 + \exp[-a(\theta - b_k)]}$$

$$= T^*(k) - T^*(k + 1) \quad ,$$

in which a is the item discrimination parameter and b_k is threshold for score-category k. $T^*(k)$ is the trace line describing the probability that a response is in category k or higher, for each value of θ. For completeness of the model-definition, we note that $T^*(1) = 1$ and

$T^*(m+1) = 0$. The value of b_{k-1} is the point on the θ-axis at which the probability passes 50% that the response is in category k or higher.

In Tests 79 and 84, there is an "item" for which the response is the test-taker's choice between problem 2 and problem 3. We treat that as an item with a binary response, $x = 1$ for the choice of problem 2 and $x = 0$ for the choice of problem 3, and use the two-parameter logistic model,

$$T(x = 1) = \frac{1}{1 + \exp[-a(\theta - b)]} \quad .$$

For the "item" in Tests 79 and 84 that represents the choice of three of five problems, we use Bock's (1972) nominal model:

$$T(x = k) = \frac{\exp[a_k\theta + c_k]}{\sum_{i=1}^{m} \exp[a_i\theta + c_i]} \quad ,$$

In this case, the data are in categories 1, 2, ..., 10, and represent the triple of problems chosen by the test-taker. There is no *a priori* ordering of these categories. The parameters denoted by a reflect the estimated order, as well as discrimination, for the categories (see Thissen, Steinberg, & Fitzpatrick, 1989). The parameters denoted by c reflect the relative frequency of the choice-categories. [Note: The parameters denoted by c in the three-parameter logistic and nominal models are not similar. However, notation for both models is established in the literature, including this conflicting use of c. So we use the established notation here; the meaning of c is clear in context.] The parameters a_k and c_k are not identified with respect to location; these parameters are estimated, subject to the constraints that $\Sigma a_k = \Sigma c_k = 0$.

III. The Results and Discussion

Test 75 was fit with the three parameter logistic IRT model (3-PL). Shown in Figure 2 is a plot of the conditional standard error of estimated proficiency [$SE(\hat{\theta} \mid \theta)$] for both test 75 and 77[5].

[5]Because both tests share the same 75 multiple choice items the underlying trait θ is essentially identical. This was confirmed when we noted that the item parameters for all of the overlapping items were identical in the two runs. This was the case for **all** of the analyses we report.

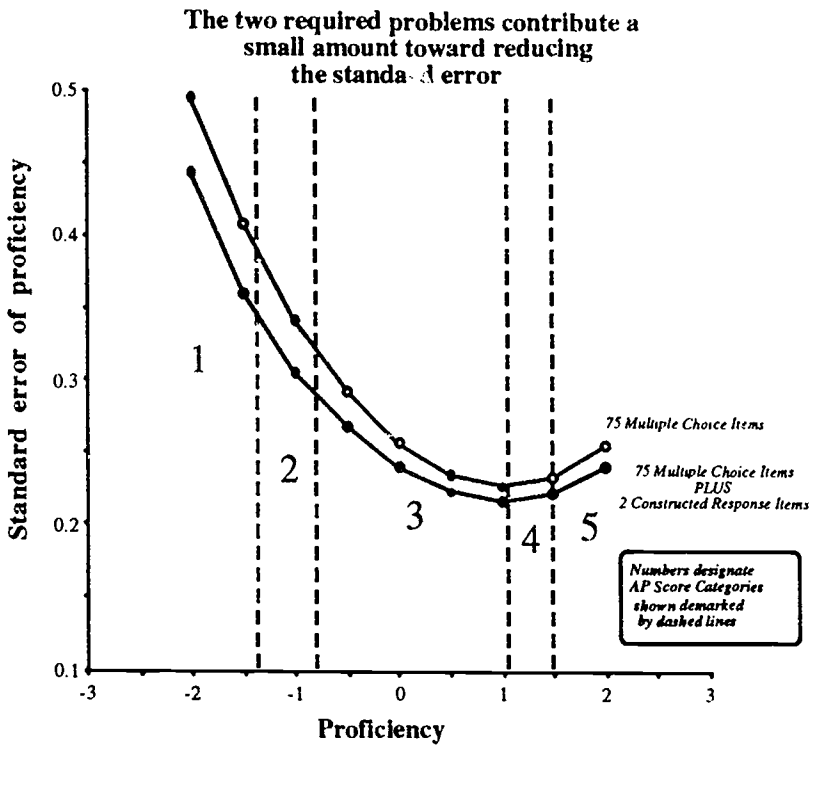

Figure 2. A comparison of the standard errors of estimate of proficiency for two versions of the AP Chemistry test. It shows that the marginal gain in precision obtained by adding two constructed-response items to the multiple choice section is modest.

This graph tells us two things.

- First that both tests yield peak information (minimum error) at the cut scores that divide category '4' students from category '3' and '5.' The cutscore dividing category '2' from '3' has a standard error about 50% larger.[6]

- Second, that the addition of two Constructed-response items (Problems 1 and 4) makes a modest contribution toward reducing the error of measurement, and that this contribution is largest at the low end of the proficiency continuum.

[6]We placed these cut scores onto this latent proficiency dimension through a procedure very much akin to equipercentile equating. Specifically, we noted what proportion of examinees were characterized by a 5. Then we calculated what value of θ corresponded to that proportion (1.5). We did this for each of the cut-scores. This procedure does not guarantee a nonlinear transformation of all raw scores onto the θ dimension , although it will be pretty close over most of the range, but it does give the right answers at the cut-scores.

To understand better the size of the contribution made by these two constructed-response items we calculated the amount of examinee time required to answer these two questions. As a rough calculation we noted that 50 minutes is allocated to Parts A, B and C, which in turn 'count' 25%, 25% and 15% respectively. The obvious calculation ($25/65 \times 50 = 19$) suggests that an examinee ought to allot 19 minutes to Problem 1 (Part A). On average an examinee can answer about 16 multiple-choice items in 19 minutes (90 minutes/75 items = 1.2 minutes/item $\Rightarrow$ 19/1.2 = 16 items). How much information is obtained from 16 multiple-choice items as compared to one constructed-response item? Shown in Figure 3 is a plot of the information from Problem 1 and from the first 16

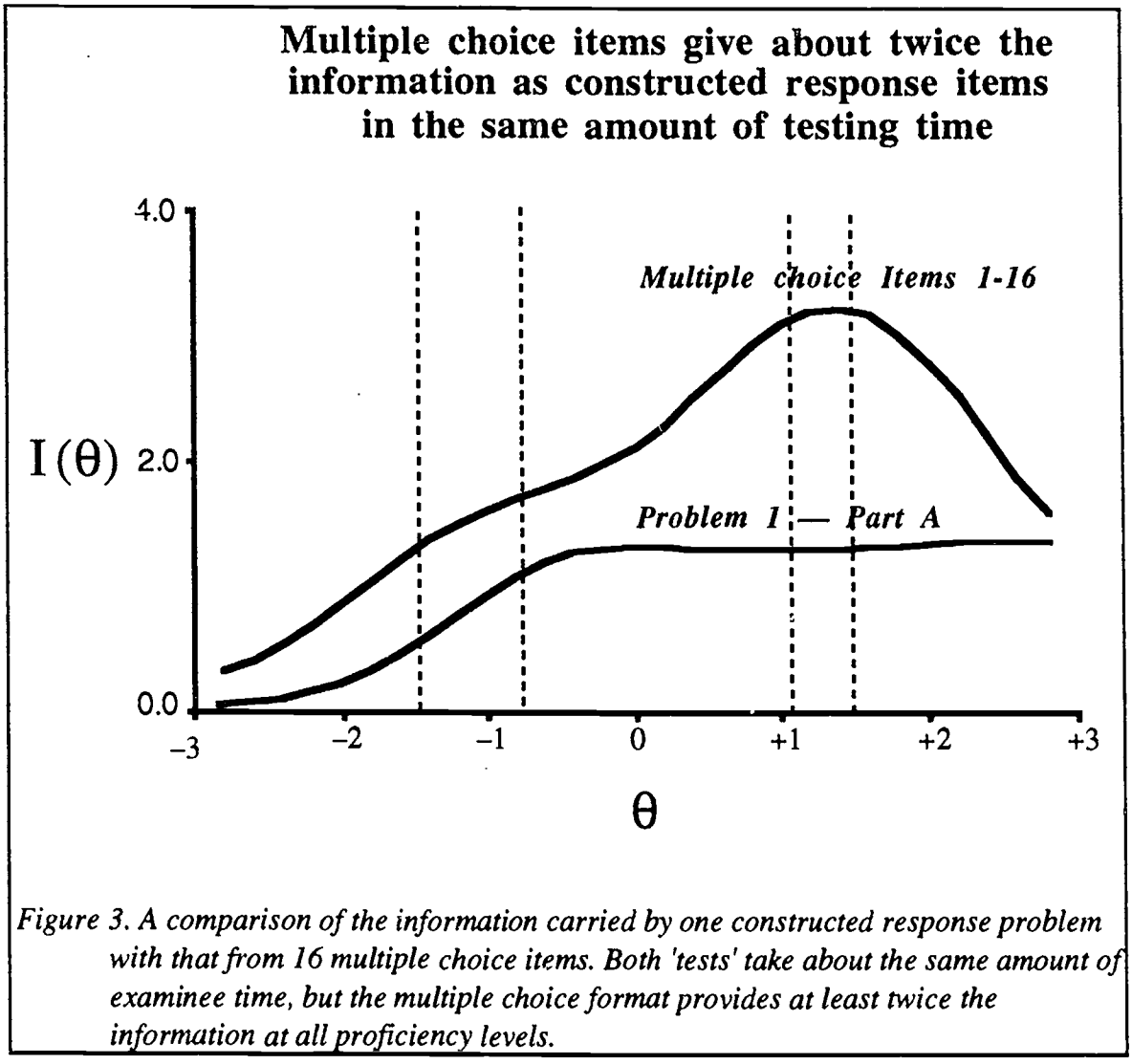

Figure 3. A comparison of the information carried by one constructed response problem with that from 16 multiple choice items. Both 'tests' take about the same amount of examinee time, but the multiple choice format provides at least twice the information at all proficiency levels.

multiple-choice items. The multiple-choice items provide more information, in the same amount of testing time, at all proficiency levels. Overall the multiple-choice items provide more than twice the information as the constructed-response item. Examining the entire test (and freely applying the Spearman-Brown prophesy formula) we found that a 75 minute multiple-choice test is as reliable as a 185 minute test built of constructed re-

sponse questions. Both kinds of items are measuring essentially the same construct (Thissen, Wainer & Wang, 1993), and the constructed-response items cost about 300 times more to score (Wainer & Thissen, 1993). It would appear, based on this limited sample of questions, that there is no good measurement reason for including constructed-response items.

We next fit Test 84. We made the untestable (and perhaps unlikely) assumption that the items not answered could be treated as ignorable and estimated proficiencies on this test. We found that there was very little additional information contained in the four additional constructed-response items added to test 77 (shown in Figure 4). What little information there was appeared at the extremes of the proficiency distribution and is unimportant for any of the decisions made with AP Chemistry Test scores.

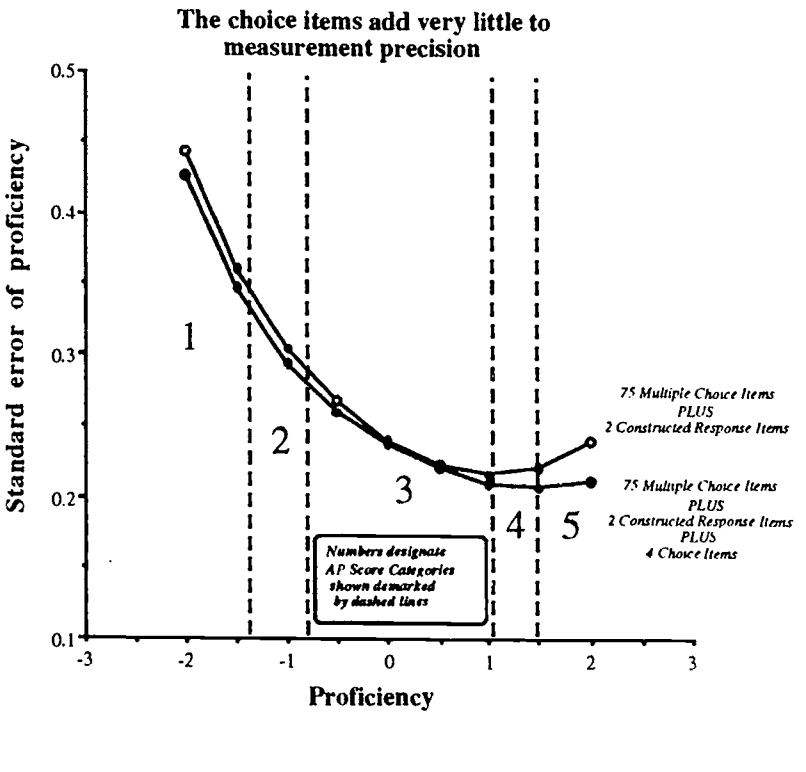

Figure 4. A comparison of the standard errors of estimate of proficiency for two versions of the chemistry test with (84) and without (77) the choice items. Even under optimistic equating assumptions, at the selection points of interest the choice items provide little additional precision.

We hasten to point out that the error function for Test 84 depicted in Figure 4 is optimistic. It is really a lower bound on the error. It attains this lower bound when the (untestable) assumption of ignorable nonresponse is true. When it is not true the error creeps upward toward the line that represents Test 77. It cannot get worse than Test 77 (identical to omitting the choice items completely) so long as the test is scored in such a

way as to weight each section by the information it contains (as was done here, within the context of an IRT model). If the test sections are weighted in some less optimal way, (i.e. specifying the number of points each item 'counts' in advance) the addition of the choice items can make the test less accurate than would have been the case had those items never been included at all.[7] This scenario is not uncommon; a number of AP tests, which include both a multiple-choice and a constructed-response section, are less reliable than their multiple choice section alone (more on this in the final section of this paper).

In an effort to obtain more information from the choice items without resorting to the possibly untenable assumption of ignorable nonresponse, we considered Test 79. We found that merely by noting which problems were chosen we were able to decrease the error of estimate at the two most important cut scores from its value for the Test 77. In addition, we found that scoring choice (as opposed to scoring the responses themselves!) yielded standard errors nearly as small as were obtained for Test 84; see Figure 5. Test 79 does not require the questionable assumptions needed to score Test 84, nor does it require the costly use of experts to score those items. In fact, it doesn't even require that the examinees use up valuable testing time to answer the items; only that the examinees indicate which items they would answer, if they were to respond.

[7]The start of the Hippocratic oath comes to mind here, "First, do no harm."

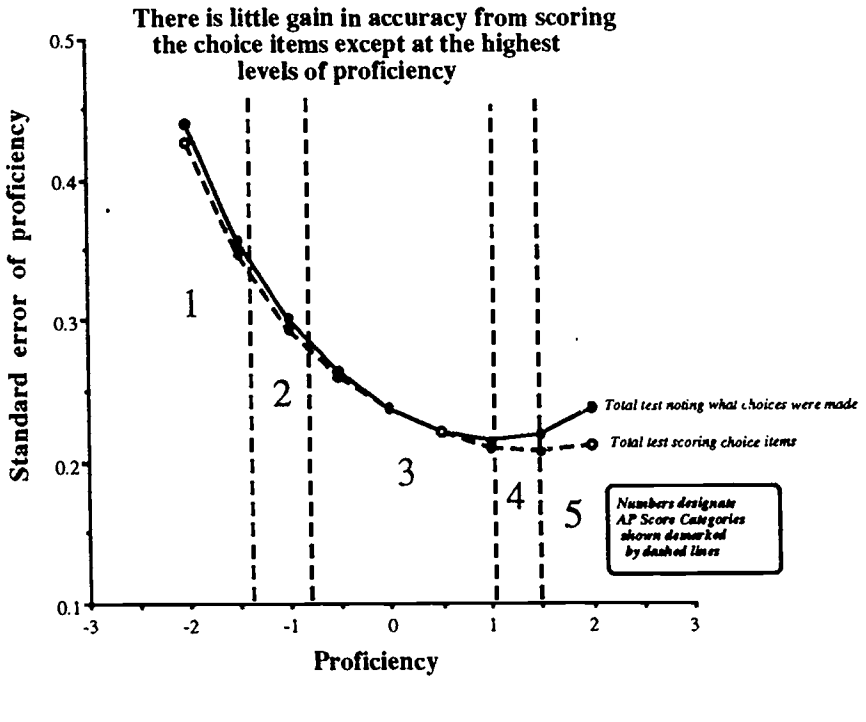

Figure 5. A comparison of the standard errors of estimate of proficiency for two versions of the chemistry test derived by scoring the choice items (84) or merely noting which items were chosen (79). At the selection points of interest scoring the choice items provides almost no practical increase in precision.

The Advanced Placement United State History Test

The 1988 AP US History test is made up of two parts. Section I consists of 100 multiple-choice questions. 75 minutes are allowed for its completion and performance on it contributes 50% of the total grade. Section II consists of two essays and 105 minutes are allotted for it. The first essay is required for all examinees. The topic for the second essay is chosen by the examinee from among five possibilities. Each essay contributes 25% toward the examinee's final score. 82,842 students took this exam.

Although we fit both sections together it is of some interest to consider results from the two sections scored separately. We fit the usual 3-PL model to the 100 multiple-choice items and obtained estimates of information from that part of the exam.

Next we fit a polytomous IRT model (Bock, 1972) to the essay portion of the test and estimated the information in that section. This fitting was done using the methodology described by Wainer, Wang & Thissen (1991)[8]. These separate results are not directly comparable to what was obtained in their joint analysis since the first analysis is information on the "multiple Choice θ" and the second on an "Essay θ" . However they are indicative of the best that each section could do when their information is measured on the trait that they were meant to measure. These results are shown in Table 1 below (as a function of a jointly estimated θ, which, as it turns out, is sufficiently like the two separate traits that we are not led astray through a comparison at this level of aggregation).

Table 1

Information in the two sections of the AP exam in United States History - 1988

	Proficiency (θ)								
	-2.0	**-1.5**	**-1.0**	**-0.5**	**0.0**	**0.5**	**1.0**	**1.5**	**2.0**
Info/section									
MC	6.2	8.0	10.0	12.0	13.7	14.6	14.3	12.8	10.5
Essay	1.0	1.0	1.0	1.0	1.0	1.0	1.1	1.1	1.1
TOTAL	7.2	9.0	11.0	13.0	14.7	15.6	15.4	13.9	11.6
Info/Item									
MC	0.1	0.1	0.1	0.1	0.1	0.2	0.1	0.1	0.1
Essay	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5
How many MC items equal one essay?	8.0	6.2	4.9	4.1	3.7	3.5	3.7	4.1	5.0
Info/minute									
MC	0.08	0.11	0.13	0.16	0.18	0.19	0.19	0.17	0.14
Essay	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01
Relative efficiency of essay items	11%	9%	7%	6%	5%	5%	5%	6%	7%

This table is divided into three sections.

The first section, comprising the top 3 lines shows the total information for the 100 MC items, for the two essays, and for the total test, if those two sections were to be combined optimally.

---

[8] The key assumption that we used to fit the model was that the tracelines for the choice items would have been the same among the examinees who didn't choose them as they were among those examinees who did. This assumption about the unobserved data is untestable but allows us to use the powerful machinery of item response theory. Also there is some experimental evidence (Wang, Wainer & Thissen, 1993) that at least in once circumstance this assumption is not too far wrong.

The second section of the table shows the average information per item. The first line is for each multiple-choice item (it comes from the first line divided by 100). The second line is for each essay (the second line of the table divided by 2). The last line is the ratio of the second line to the first. This line gives a rough estimate of how many multiple-choice items one essay is worth in terms of information. The answer ranges from 4 to 8 depending on the proficiency of the examinee.

The last section of the table examines the amount of information obtained per minute of testing time. The 100 multiple-choice items are administered in 75 minutes, and so the figures shown here are derived by dividing the first row of the table by 75. The next line in the table is the information per minute for the essay portion. The essays are administered in 105 minutes and so this line is derived by dividing the second line of the table by 105. The last line is the ratio of these two and can be interpreted as the relative efficiency of the essay section of the test when compared to the multiple-choice portion. An average essay's efficiency ranges from 5 to 11 percent of that of an average multiple-choice item per minute depending on the proficiency of the examinee. Thus for a middle level examinee (θ=0) one would need about 20 times as much examination time to get the same amount of information with an essay as one would obtain with multiple-choice items.

IV. Conclusions and Implications

The analyses reported here have implications in three separate, but practically related, areas. The first is on the value of constructed-response items. The second on the gains and losses associated with allowing choice. The third, on the practicality of complex IRT models for scoring and analyzing operational tests. We will discuss each of these in turn.

On the value of constructed-response items

On the basis of the data we examined we are forced to conclude that **constructed-response items provide less information in more time at greater cost than do multiple-choice items.** This conclusion is surely discouraging to those who feel that constructed-response items are more authentic, and hence in some sense, more valid than multiple-choice items. It should be.

There is no evidence to indicate that these two kinds of questions are measuring fundamentally different things, at least for domains such as those represented by AP Chemistry, US History or Computer Science. One interpretation of the results obtained from earlier dimensionality studies (Thissen, Wainer & Wang, 1993) is that there are two underlying dimensions in the AP Chemistry test, but that there is very little information in the data about a factor specific to the constructed-response items. Another interpretation

of those factor analytic results is that the there are two highly correlated dimensions, with substantial loadings of all items on both. One dimension is slightly more related to the multiple-choice items , while the other is slightly more related to the constructed-response items. Thus if we are interested in accurately measuring the "constructed-response proficiency" we could do so almost as well using the multiple-choice items as we could using the constructed-response items. Such a measure would be more accurate than a constructed-response test taking the same amount of examinee time because the statistical bias introduced by measuring the wrong thing (the "multiple-choice proficiency") is small in comparison with error variance associated with constructed-response items. This interpretation is supported by the results we obtained from the US History test as well as correlations on other AP tests (reported in Table 2).

We need to emphasize here that the constructed-response items in the AP Chemistry test are high quality items. They are scored using an analytic scoring scheme that minimizes the inter-rater variability. The advantage enjoyed by multiple-choice items increases on tests in which the constructed-response items are scored holistically, and hence have much lower inter-rater reliability [the AP exam in *Music: Listening and Literature* (no longer being given) would require more than 10 hours of constructed response items to get the same precision that is obtained from 15 minutes of multiple-choice testing (Wainer & Thissen, 1993)]. Last, because of practical limits on testing time, topic coverage using constructed-response items may be more limited. This has prompted the testing strategy of allowing examinee choice. We have shown here that the choice items given in Chemistry, under the most benign of circumstances (ignorable nonresponse) provide very little extra information, almost none at the most important points on the proficiency distribution. But, if that assumption is not true (although this is untestable with the data on hand, it is a likely possibility) allowing choice can make the test worse.

These results, obtained as they have been from two widely disparate kinds of tests, tell much the same story. How generalizable is this tale? We believe that all of the Advanced Placement tests would show the same effect to some extent. Exactly how much less efficient the constructed-response items are, relative to multiple-choice items, awaits formal analyses parallel to these. It appears that holistically scored items do worse than analytically scored ones. We conjecture that the closer an item gets to being scored objectively the smaller the difference in efficiency we would observe between that item and parallel items in a multiple-choice format. We base these conjectures about generality and relative efficiency on the analyses we have done as well as on strong hints seen as part of *de facto* analyses done on all AP tests. For example, consider the results shown in Table 2 on seven Advanced Placement tests aggregated over a five year period. We see that in all cases the multiple-choice portion of the test correlates more highly with the constructed-response portion than the CR portion does with itself (its reliability). This means that if we wish to predict a particular student's score on a future test made up of constructed-response items we could do so more accurately from a multiple-choice than

from a constructed-response test. Thus, while we are sympathetic to the sorts of arguments we cited at the beginning of this paper regarding the advantages of a constructed-response format, we have yet to see convincing evidence supporting them. We are awash in evidence of their drawbacks. Aside from their increased costs, both in scoring and examinee time, note that in five of the seven cases the inclusion of the constructed-response portion actually lowered the test's reliability. Again, the beginning of the Hippocratic oath comes to mind.

Table 2

Advanced Placement Statistics - Five year medians 1982-1986

Test	Reliability		Correlation MC & CR	Reliability of Composite Score
	Multiple Choice	Constructed Response		
Mathematics: Calculus AB	0.89	0.80	0.83	0.92
Computer Science	0.88	0.79	0.80	0.90
Chemistry	0.91	0.77	0.84	0.90
French Language	0.93	0.75	0.78	0.92
Biology	0.93	0.68	0.73	0.89
European History	0.90	0.46	0.50	0.80
Music: Listening & Lit.	0.85	0.29	0.47	0.84

Data source: *Technical Manual for the Advanced Placement Program for* 1982-1986. College Entrance Examination Board: New York, 1988.

The generality of these findings may be limited. It might be argued that the constructed-response items used in these tests are not state-of-the-art. Or that the scorers are not as well trained as they might be. Or that the scoring rubrics are deficient in some important way. While all of these excuses may be justified, we are not sanguine about their easy amelioration. In fact, these test items and their scoring rubrics are built with the care and expertise borne of many years of experience. The judges are trained and checked using procedures that have developed and grown with that experience. While improvements are possible in many areas, it does not seem likely that these improvements would yield a radical change in the result. At least not for an operational test with thousands of examinees.

It is instructive to note that a recent report (Koretz, et al, 1992) on the reliability of the 1992 Vermont Portfolio Assessment Program found that despite "Vermont educators" finding "the program burdensome" (p. 1) "the overall pattern (of the assessment) was one of low reliability" (p.2). The average reliabilities ranged, depending on subject and grade

level, from .33 to .43. These do not compare favorably with the constructed-response reliabilities for AP tests given in Table 2.

On some gains and losses associated with allowing choice

An important difference between multiple-choice and constructed-response questions is the time it takes to complete a test item. A typical multiple-choice test may allow about one-half to one minute per question, perhaps slightly more if long reading passages or complex mathematical computations are involved. Other formats, such as multistep computations, diagrams, and/or explanations, may be scheduled to take 10 minutes each; while essays may require a half an hour or more per question.

No test can completely cover all possible topics in a subject at all levels of difficulty, and hence a test developer must sample from the domain of knowledge that the test is designed to represent, and try to be as comprehensive as possible in the time available. The allocation of scarce time resources becomes especially acute with constructed-response test questions because so many fewer can be given in even a three- or four-hour test. The trade-off between breadth of coverage and practical time limitations has led test developers to offer examinees a choice among items. The evidence available from the choice items in the AP testing program suggests that all examinees do not choose wisely, and that choice wisdom varies with the sex and ethnicity of the examinee.

Support for the first observation was provided from a special data gathering in which it was discovered that a substantial number of examinees would have received a higher score on the item that they did not choose than they received on the one they did (Wang, Wainer & Thissen, 1993).

The second observation may be inferred from existing data sources. For example, we can measure the extent to which allowing examinees to choose among items provides fair measurement with the same tools that are used to assess item fairness in other contexts. In this spirit we examined the choices made in Part D of the chemistry test for sex DIF. The 'item' is "Pick 3 items from the set (5,6,7,8,9)." Using the methodology described in Wainer, Sireci & Thissen (1991) we fit an anchor test consisting of 10 multiple choice items and the two required constructed-response items (problems 1 and 4) and then estimated the fit of this model when the Part D choice problem was included and forced to have the same parameters for males and females, and then again when it was allowed to have different parameters in those two groups. The difference in the -2loglikelihoods for these two models was 360 with 18 degrees of freedom. This is thus a χ^2 test of the hypothesis that, conditioned on proficiency, men and women make the same choices. Obviously this hypothesis can be rejected. This finding provides yet another reason why it is important to equate choice items for differences in difficulty.

One needn't resort to complex analyses to see this effect. For example in Figure 6 is shown the proportion of examinees who chose each essay in the US History test as a function of sex and performance on that essay. Essay 3 had the lowest average scores for both men and women. The usual interpretation of this finding is that the topic of Essay 3 is the most difficult. This topic was about twice as popular among women as among men.

Both of these results provide strong evidence that if choice items are not equated for their differential difficulty sex differences will be artificially exacerbated. But if they can be successfully equated why did we need to allow choice in the first place?

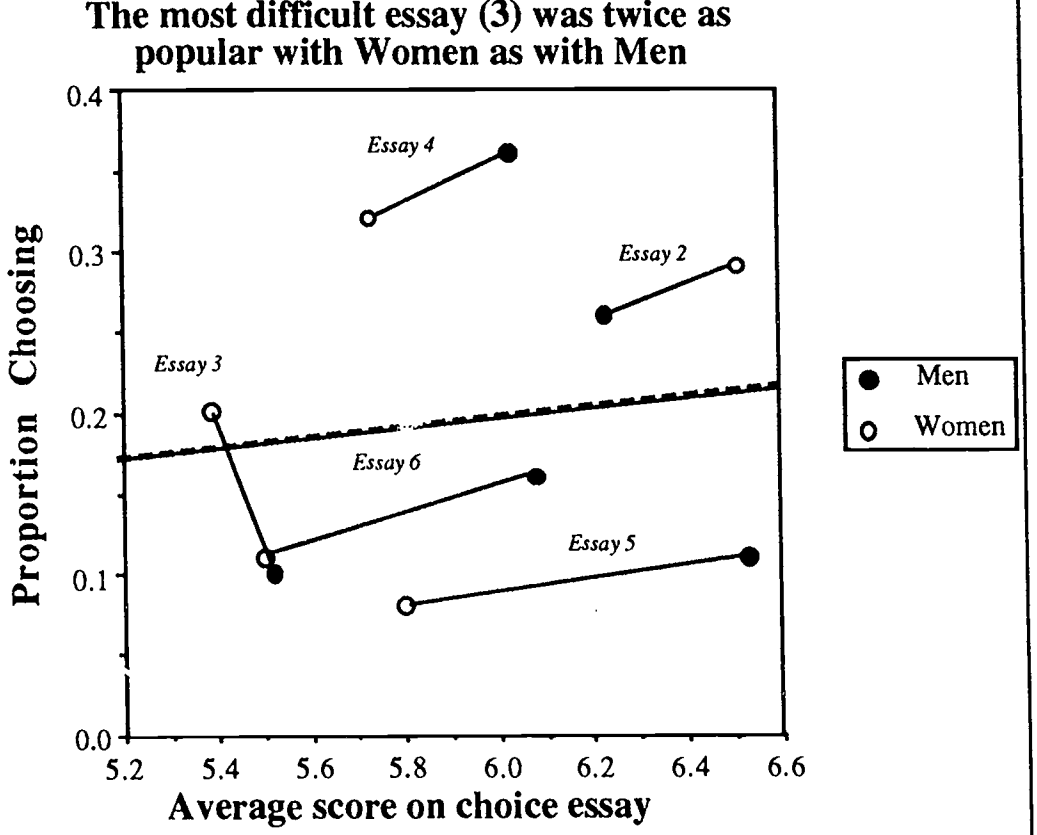

Figure 6. A summary of male and female examinees' performance on the choice essays of the 1988 AP History test. Essay 3 generates the lowest scores of all choices and is selected by twice as many women as men. All observed differences are statistically significant.

On the practicality of complex IRT models for scoring and analyzing operational tests

The AP Chemistry test has a total of 84 items of wildly varying item types. More than 18,000 examinees took it. The United States History test contains 106 items of two different formats and more than 80 thousand examinees. We scored both of them using the program MULTILOG on a 386 PC. An average run might have taken 7 or 8 hours. While such processing times don't encourage computationally intensive methods (such as multiple imputation to study the sensitivity of the results with respect to missing data), they are not impractical for scoring an operational test. One merely prepares the data during the day and sets it running before going home for the evening. In the morning, if there have been no intervening power failures, the results are waiting for you. Of course faster machines with more memory will yield shorter turn-around times. Specialized computer programs optimized for the task would help even more.

We would like to draw the reader's attention to the complexity of the models used in the analysis of "Test 84" in Chemistry. It is a mixture of IRT models and contains three of thepossible characteristics: binary, nominal, and ordered category . This is the first application we know of that has used all of MULTILOG's models together. An earlier application (Roche, Wainer & Thissen, 1975) used such a complex mixture to measure skeletal age, but the methodology used was sequential and sometimes more than a little *ad hoc* because neither MULTILOG nor maximum marginal likelihood estimation had yet been invented. Consequentially the solution arrived at in that early application did not enjoy the joint optimization that is yielded by the coherent model fitting described here.

What is to be gained through the use of such computationally complex models for test scoring? In a phrase, improved measurement precision and improved knowledge of what is that precision. To illustrate how this might work, let us return to the error functions shown in Figures 2, 4 and 5. All four of the tests analyzed show essentially the same structure. To wit, that the error of estimate of proficiency is about 50% higher at the 2-3 decision point than at the 3-4 decision point.[9] The small amount of information yielded by constructed-response problems seems more focused at the high end of the proficiency distribution. Yet the most important decisions are made at the 2-3 cut-point, for (according to the instructions attached to the released test form, p. 30) a score of '3' is defined as "qualified" for advanced placement. This observation suggests that the test is somewhat mis-aimed and is too difficult for the decision task for which it was built. Can any post-hoc scoring scheme help?

---

[9]We placed these cut scores onto this latent proficiency dimension through a procedure very much akin to equipercentile equating. Specifically, we noted what proportion of examinees were characterized by a 5. Then we calculated what value of θ corresponded to that proportion (1.5). We did this for each of the cut-scores. This procedure does not guarantee a nonlinear transformation of all raw scores on the θ dimension, although it will be pretty close over most of the range, but it does give the right answers at the cut-scores.

Previous research (Bock, 1972; Thissen, 1976) has suggested that there is information in noting which wrong answers are chosen. Of course this lode of information is richest where wrong answers are most plentiful; the lower portions of the proficiency distribution. But happily, this is exactly where we need more precision. Moreover, technology has improved substantially in the past two decades to make it possible (indeed practical) to fit a polytomous nominal IRT to a 75 item 5-choice test (Ramsay, 1991; 1992). When this was done the error functions shown in Figure 7 obtained. Thus the 75 item multiple choice test when scored polytomously yielded a test that dominates, in accuracy, the optimally (and optimistically) scored full 84 item test at all of the choice points of importance. It also corrects the problems associated with the difficulty of the original test being too high.

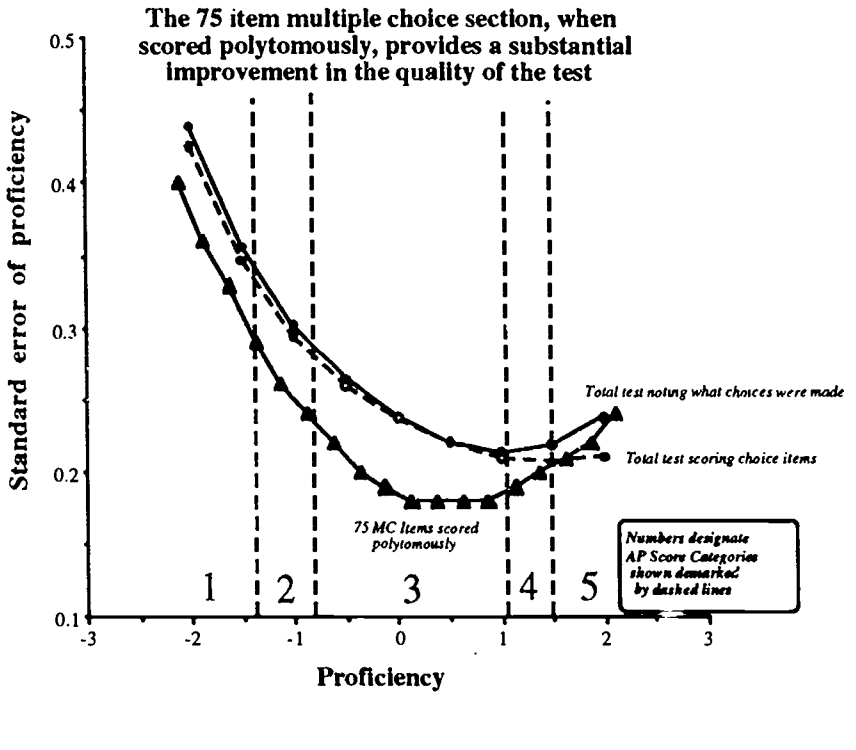

Figure 7. The point of maximal precision of measurement for the AP Chemistry test is higher then the most important choice points. Matters can be improved considerably using only the multiple choice portion of the test and utilizing the information in examinees' wrong responses.

References

Bennett, R. E., Ward, W.C., Rock, D. A., & Lahart, C. (1990). *Toward a framework for constructed-response items* (RR-90-7). Princeton, NJ: Educational Testing Service.

Bennett, R. E. (1991). Toward intelligent assessment: An integration of constructed-response testing, artificial intelligence, and model-based measurement. In N. Frederiksen, R. J. Mislevy, and I. Bejar (Eds.), *Test theory for a new generation of tests*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Birenbaum, M., & Tatsuoka, K. K. (1987). Open-ended versus multiple-choice response formats — It does make a difference for diagnostic purposes. *Applied Psychological Measurement, 11,* 385-395.

Bock, R. D. (1972). Estimating item parameters and latent ability when responses are scored in two or more latent categories. *Psychometrika, 37,* 29-51.

Bock, R.D., & Aitkin, M. (1981). Marginal maximum likelihood estimation of item parameters: An application of an EM algorithm. *Psychometrika, 46,* 443-459.

Fiske, E. B. (1990, January 31). But is the child learning? Schools trying new tests. The *New York Times*, pp. 1, B6.

Frederiksen, J. R., & Collins, A. (1989). A system approach to educational testing. *Educational Researcher,* 18(9), 27-32.

Glaser, R. (1985). The integration of instruction and testing. A paper presented at the ETS invitational Conference on the Redesign of Testing for the 21st century, New York, New York.

Guthrie, J. T. (1984). Testing higher level skills. *Journal of Reading, 28,* 188-190.

Koretz, D., McCaffrey, D., Klein, S., Bell, R., & Stecher, B. (1992). *The reliability of scores from the 1992 Vermont Portfolio Assessment Program.* Interim Technical Report. Santa Monica, CA: Rand Institute on Education and Training.

Little, R. J. A., & Rubin, D. B. (1987). *Statistical analysis with missing data.* New York: Wiley.

Lord, F. M. (1980). *Applications of item response theory to practical testing problems.* Hillsdale, NJ: Lawrence Erlbaum Associates.

Lord, F. M., & Novick, M. (1968). *Statistical theories of mental test scores*. Reading, Mass.: Addison Wesley.

Nickerson, R. S. (1989). New directions in educational assessment. *Educational Researcher*, 18(9), 3-7.

Pollack, J., Rock, D., & Jenkins, F. (1992). *Advantages and disadvantages of constructed-response item formats.* Paper presented at the annual meeting of the AERA, San Francisco, CA.

Ramsay, J. O. (1991). Kernel smoothing approaches to nonparametric item characteristic curve estimation. *Psychometrika, 56,* 611-630.

Ramsay, J. O. (1992). *TESTGRAF: A program for the graphical analysis of multiple-choice test and questionnaire data.* Technical Report . Montreal, Quebec: McGill University.

Roche, A.F., Wainer H., & Thissen, D. *Skeletal maturity: The knee joint as a biological indicator.* New York: Plenum, 1975.

Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. *Psychometrika Monographs,* (Whole No. 17).

Sebrechts, M. M., Bennett, R. E., & Rock, D.A. (1991). *Machine-Scorable Complex Constructed-Response Quantitative items: Agreement Between Expert System and Human Raters' Scores.* ETS Research Report 91-11. Princeton, NJ: Educational Testing Service.

Tatsuoka, K. K. (1991, August). *Item construction and psychometric models appropriate for constructed-responses.* ETS Research Report. Princeton, NJ: Educational Testing Service.

Thissen, D. (1976). Information in wrong responses to the Raven Progressive Matrices. *Journal of Educational Measurement,* 13, 201-214.

Thissen, D. (1982). Marginal maximum likelihood estimation for the one-parameter logistic model. *Psychometrika, 47,* 201-214.

Thissen, D. (1991). *MULTILOG user's guide* (Version 6). Chicago, IL: Scientific Software.

Thissen, D., & Steinberg, L. (1984). A response model for multiple-choice items, *Psychometrika, 49,* 501-519.

Thissen, D., Steinberg, L., & Fitzpatrick, A.R. (1989). Multiple-choice models: The distractors are also part of the item. *Journal of Educational Measurement, 26,* 161-176.

Thissen, D., Wainer, H. & Wang, X-B. (1993). *How unidimensional are tests comprising both Multiple-choice and Free-Response Items? An analysis of two tests.* ETS Technical Report (93-xx). Princeton, N.J.: Educational Testing Service.

Wainer, H., & Thissen, D. (1993). Combining multiple-choice and constructed-response test scores: Toward a Marxist theory of test construction. *Applied Measurement in Education, 6,* xxx-xxx.

Wainer, H., Sireci, S.G. & Thissen, D. (1991). Differential testlet functioning: Definitions and detection. *Journal of Educational Measurement, 28,* 197-219.

Wainer, H., Wang, X. B., & Thissen, D. (1991). *How well can we equate test forms that are constructed by examinees?* Technical Report (91-15). Princeton, N.J.: Educational Testing Service.

Wang, X-B., Wainer, H. & Thissen, D. (1993). *On the viability of some untestable assumptions in equating exams that allow examinee choice*. Technical Report (93-xx). Princeton, N.J.: Educational Testing Service.

Yamamoto, K., & Kulick, E. (1992). *An information-based approach to monitoring content validity and determining the relative value of polytomous and dichotomous items.* Paper presented at the annual meeting of the AERA, San Francisco, CA..