MANAGING THE INFLUENCE OF DIF FROM BIG ITEMS: THE 1988 ADVANCED PLACEMENT HISTORY TEST AS AN EXAMPLE

Howard Wainer
Robert Lukhele

Educational Testing Service
Princeton, New Jersey
May 1996
Managing the influence of DIF from big items: The 1988 Advanced Placement History test as an example

*Howard Wainer & Robert Lukhele*

*Educational Testing Service*

Abstract

Building tests out of large items brings with it a number of problems. One major problem is that it is often too difficult and too expensive to extensively pretest large items. Thus the sorts of screening for flaws that are *pro forma* for multiple choice items is not often done for large items. In addition, since there are so few large items on an operational test, not counting an entire item that is found to be flawed in an operational administration may be tantamount to aborting that administration. In this paper we examine the efficacy of the alternative of continuous item weighting. This alternative is illustrated on data from the 1988 administration of the College Board's Advanced Placement History Test.

This article will be appearing in

*Applied Measurement in Education, 10(2), 1997*

all references should be to that source
Managing the influence of DIF from big items:
The 1988 Advanced Placement History test as an example

Introduction

Differential Item Functioning (DIF) is a term that has been adopted to characterize the extent to which there are idiosyncratic group differences in performance on an item. "Idiosyncratic" in this instance means that the group differences are observed between comparable strata of examinees. All things being equal, contemporary test construction practices prefer, if possible, to not include items with appreciable DIF on any operational test forms. This goal is commonly achieved through extensive pretesting of items. When possible, substitutes are found for any items discovered in pretesting, to show DIF with respect to any of the standard focal groups. The screening effects of pretesting are usually sufficient to remove virtually all DIF-laden items from operational tests (Burton & Burton, 1993). In the rare event of an item with DIF appearing on an operational test one of the actions available is to elide the item from the test post hoc and provide a score without it.

The procedures for dealing with DIF just described are not well suited for large\(^1\) test items (essays, large problems, performance tasks, portfolio components). For both practical\(^2\) and security\(^3\) reasons such items are rarely pretested. Because of these limitations such pretests that are carried out tend to be on smallish samples which do not have a large enough representation of many of the focal groups of interest to yield nominally accepted levels of statistical power in the detection of DIF. Without pretesting the only other option so far considered for dealing with a large item that shows DIF in operation is to eliminate it from scoring. This option is undesirable for at least three reasons. It would be:
1. Wasteful of money - it is usually very expensive to score a large item. Who, after spending hundreds of thousands of dollars scoring an essay, would then be willing to discard it?

2. Wasteful of examinee effort - if examinees spend 30 minutes of a 2 hour exam writing an essay, how can we justify telling them that we will not be counting it?

3. Wasteful of information- discarding a substantial sample of examinee effort reduces the information in the test score. In doing so, it may reduce the test's reliability and validity below what are acceptable minimums and require nullifying the entire administration.

What other alternatives are there for dealing with DIF? A possibly useful perspective in the search for alternatives is to note that current policies essentially attach a weight to every item in the calculation of an examinee's score that corresponds to the item's DIF. If there is enough DIF, the weight is zero; if DIF is smaller than that, the weight is one. It is this all-or-none approach that leads to the waste described previously. Perhaps by taking an intermediate path we can reduce the influence of an item's DIF on final score sufficiently to make it acceptable, while at the same time allowing us to harvest some of the valid information that it contains. In the balance of this paper we discuss the value of a scheme that weights each item by the size of its relationship to the underlying trait being measured. We then illustrate how it works on data from a 1988 administration of the College Board's Advanced Placement US History Test.

The method that we propose here is only meant to be applied after a careful examination of the DIF-laden item reveals neither offensive content nor evidence that its content is outside of the intended domain of the test. In these cases we believe that eliding the item completely is likely to be the proper strategy. Next, we do not intend that the strategy we propose be implemented just for large items. This would mean a different
standard of acceptance for large items than for small. While there are practical reasons for having this (eliding a small item is a much more practical option), it isn't necessary. We believe that this same strategy, differential DIF weighting, can be applied for all items types and so yield a completely consistent strategy. Finally, the IRT strategy proposed next is only suitable when there are some other data available that allows the examinees' performance without the studied item to be stratified. In the case we discuss it is the rest of the test. Other stratifying variables are possible.

One Solution

Item weighting

The contribution of a large item to an examinee's final score is usually determined a priori and is roughly related to the amount of time allocated for that item's completion (e.g., an essay that takes 25% of the testing time usually 'counts' for about 25% of the grade). It has been shown (Lukhele, Thissen & Wainer, 1994) that large items usually contain far less information than an equivalent (in terms of examinee testing time) number of multiple choice items. For the 1988 AP US History Test the essay portion is overweighed by (roughly) a factor of twenty. This may be discouraging for proponents of largish items, but may save the day if a large item is found to contain substantial DIF. If we weight items by their informational contribution rather than some a priori scheme two outcomes should ensue. First the reliability of the test will increase. Second, if a large item does contain DIF, the effect of that DIF on an examinee's score will be reduced by (in this instance) as much as a factor of twenty. In some cases this may make the size of the DIF sufficiently small to be acceptable. The canny reader will already discern that one such weighting scheme is nothing other than that which occurs naturally from IRT. Indeed, that is the approach that we will be pursuing shortly.
Why do we expect this to work? There is one principal reason. As already mentioned, big items tend to be overweighed in estimates of their empirical value to the overall score. This is because most big items are scored by expert judges. Hence they have one additional source of error that is not found in objectively scored (i.e., multiple choice) items. Thus ceteris paribus, the scoring method alone will yield items containing greater measurement error. But all things are not equal. In addition to the greater scoring error, big items usually tend to take more examinee time to answer. Thus there are at least two facets to the advantage multiple choice items have in any measure of information per examinee minute. Thus by weighting large items by their informational content we improve the test's performance while simultaneously decreasing any DIF-related effects of the item.

An example: 1988 AP US History three ways

The Advanced Placement United States History (AP US History) examination is divided into two sections (Section I and Section II) with 75 minutes allocated for Section I and 105 minutes allotted for Section II. Of the 105 minutes, 15 minutes are for reading the questions and 90 minutes for writing the two essays.

Section I consists of 100 five-option multiple-choice (MC) questions. Each correct MC item is worth 1 point, an incorrect MC item is worth zero. This section accounts for 50% of the total grade.

Section II consists of essay questions, and has two parts:

1. Part A consists of document based question (DBQ), Question 1, that all examinees must answer. It is scored on a 16 point scale (0-15) and accounts for 25% of the total grade. 40 minutes are allocated for this question.
2. Part B has five essays (Questions 2 through 6), and the examinee must answer exactly one of those. It is scored on a 16 point scale (0-15) and accounts for 25% of the total grade. 50 minutes are allocated for this question.

The composite score on the AP US History test is obtained through a weighted combination of the observed scores on the multiple choice (MC) section, the document based question (DBQ), and the essay whose topic is chosen by the examinee from the five offered.

Composite Score (180 points) = 0.9 (MC total) + 3 (DBQ) + 3 (Choice Essay)

<table>
<thead>
<tr>
<th>Points</th>
<th>Possible</th>
</tr>
</thead>
<tbody>
<tr>
<td>90</td>
<td>45</td>
</tr>
</tbody>
</table>

This form of the examination was taken by 82,978 examinees in 1988. The test form has been released.

Shown in Table 1 are the mean scores for men and women on the five choice essays (labeled 2 through 6). While these unadjusted means do not, in themselves indicate DIF, under an assumption of random choice⁵ essay 5 offers the most likely place to find DIF if there is any. Thus for this illustration we confined our investigation to a random sample of 2,000 of those examinees who chose essay 5. This sample consisted of 1,244 men and 756 women, an amount that is more than ample for stable estimates of all parameters of interest.

Outline of the analyses

The goal of this example is to illustrate how we can reduce the effect that a differentially functioning large item has on total test score without adversely affecting the pre-
cision of the test. We will accomplish this by reducing the contribution that the offending item makes toward total score. The effect that we can demonstrate using real data from an operational test will be very small because the care that is lavished on test construction is generally rewarded with high quality test items. Finding an item with any DIF is hard. Finding one with DIF large enough to yield a dramatic example is next to impossible. We hope that the technology we are proposing will not often be required, but we believe that it is important to have on hand for future emergencies, no matter how unlikely. The demonstration has four principal steps

**Step 1. Fit the entire test with a hybrid IRT model allowing for DIF**

Fit a hybrid IRT model (3-PL for 100 MC items and a graded polytomous model for the two essays) that adjusts for any sex DIF that might exist in essay 5 following the procedures described by Wainer, Sireci & Thissen (1991) which estimate parameters for the item separately by sex. Each item is optimally weighted by its relationship to the underlying trait. This approach, though in some senses psychometrically optimal, is illegitimate in practice; adjusting for sex DIF means adding (or subtracting) some amount to an examinee's score based on their sex. But this provides a benchmark for the distributions of ability for us to compare other scoring schemes. In a very real sense this is the best estimate of the distributions of proficiency by sex.

**Step 2. Redo the IRT analysis in step 1, but without adjusting for DIF**

The difference between the fit statistics obtained from these two analyses (characterized by $-2 \times \text{loglikelihood}$) provides a statistical test of the significance of the DIF observed. If there is no DIF we have no example to illustrate our advice. Obviously (since you are reading this) there did turn out to be a significant, albeit small, amount of sex DIF on essay 5. The estimates of proficiency obtained from this analysis will be affected by the DIF in essay 5.
Step 3. Score the test in the traditional way

Each item is weighed by the amounts specified in the exam book and reported in the previous section.

Step 4. Transform the IRT proficiency estimates into the observed score metric

This is the crucial step to assess the efficacy of the approach we are espousing here. Using the correspondence table (provided in the technical documentation for the test) between the raw scores on the test, the boundaries of the five score categories, and the percentage of examinees within each category we are able to equate the IRT proficiency scale to the raw score scale. We then compare the mean scores (in the observed score metric) for each sex group in each of the three analyses. The size of the difference between the results of the traditional method and those from method one characterizes the size of the problem. The score estimates from the second analysis should lie between those from the first and the third. The closer they are to those from the first analysis the better. The smaller the difference between the first and second analyses the more that the effect of DIF on an essay can be dissipated solely through optimal weighting.

All analyses were done using MULTILOG Version 6.0 (Thissen, 1990); it allows the mixing of item types within the same analysis. Such mixing is crucial in situations like that in Analysis 1 in which we use an external anchor of 100 dichotomous items and one polytomous item. It also allows the imposition of equality constraints; that is necessary to obtain the likelihood of restricted (no DIF) models (Analysis 2).
Analysis 1 - Full IRT with Sex DIF

The 100 multiple choice items were fit with the traditional three parameter logistic (3-PL) IRT model (Lord, 1980) using standard default priors on the parameters during estimation. The two essays were fit using Samejima's (1969) graded model. These models are rather well known, and so we shall not repeat their description here. Readers unfamiliar with these details are referred to the original sources or, for a compact description, to Wainer (1995, especially pages 166-167).

The hybrid mixture of Samejima's graded model with the three parameter logistic model provides a good representation of the data. All items were fit simultaneously, which lends a coherence to the analysis that is difficult with outmoded methods in which each item type is fit separately and then joined through some sort of connecting transformation.

We felt justified in using a single trait to characterize performance for two reasons. First, because no attempt is made on any AP exam to provide anything other than a single score for each examinee, we were mirroring operational practice. The single score provided by IRT differs from that given in operational practice only in that it has at its base some psychometric optimality characteristics that are lacking in the current scoring procedure. The second reason is statistical; many recent factor analytic studies using the most contemporary methods (Bock, Gibbons, & Muraki, 1988) repeatedly show that the unidimensionality assumption is a close enough approximation to allow IRT to provide a more than adequate fit to the data in all Advanced Placement tests examined (Bennett, Rock & M. Wang, 1991; Thissen, Wainer, & Wang, 1994). Their results led Thissen et al (1994, p. 120) to conclude that, "there is also clear evidence that the free-response problems predominantly measure the same thing as the multiple-choice sections."
We found further support for this with these data. While the reliability of the total test was .83, the reliability of the multiple choice section of the test was .90; the free response section was .51. Obviously if the goal was to maximize reliability one should down weight the free response section substantially. Interestingly, the multiple choice section correlated .54 with the free response section. One interpretation of this is that we can get a modestly better estimate of the proficiency measured by the free response section from the multiple choice score than from the free response score. This finding is entirely consonant with results from many other hybrid tests.

In this analysis we treated essay 5 as two separate essays; we acted as if one (denoted 5*) was administered only to males and the other (5**) was administered only to females. The structure of the analysis is shown in figure 1. Shading indicates that the indicated group was presented with the item. This artificial splitting up of Essay 5 is a common computing trick that allows us to estimate parameters for that essay separately for males and females while using all the rest of the test as an anchor. This assures that the parameters thus estimated will be on the same scale. The scale was set by fixing the male proficiency distribution as normal with mean zero and standard deviation one [N(0,1)]. All estimates of parameters were obtained by maximizing their marginal likelihood (Bock & Aitken, 1981).

-------------------------------------------------------------
Insert Figure 1 about here
-------------------------------------------------------------

Results of fitting:

After fitting this general model to the data we obtained a -2loglikelihood of 960,734. This number is of virtually no interest by itself, but when subtracted from its
counterpart obtained in the second analysis, it will provide strong evidence that this item contains significant sex DIF.

It is important to transform this IRT analysis into an observed score metric to provide a common metric of comparison with the observed score approach described in the third analysis. Thus after estimating proficiency with this model we transformed the IRT scale \([N(0,1)]\) into the same metric as the observed scores through an equipercentile equating at the four points in the reported metric. That is we match the value of \(\theta\) that cuts off the same percentage of examinees as the reported raw score for each of the reporting categories. The results of that equating for the situation in which proficiency is estimated without regard to sex are shown in Table 2 below.

=================================
Insert Table 2 about here
=================================

Thus the value of \(\theta\) corresponding to the top 10.3\% is 1.31 which corresponds to the raw score of 117. After continuing in an identical fashion for the other three points we obtained four points of correspondence between the two scales. We then fit an interpolating polynomial between these points for intermediate values and bent over the ends to match the extrema. Although an equipercentile equating with this sparseness of anchor points is rather rough, it is more than adequate for the task at hand here.

We found that after adjustment for DIF the male mean was 10.5 points higher than the female mean (male = 87.1; female = 76.6). This is our best estimate of the actual difference in the proficiencies of each sex. The distributions of scores, by sex are shown graphically in Figure 2 below.
Analysis 2 - Full IRT with No Sex DIF

Analysis 1 involved fitting a completely unrestricted model—estimating all of the parameters separately for both the reference and the focal groups -- and noting the value of -2loglikelihood (asymptotically $\chi^2$) for that model. In Analysis 2 we restrict the parameters for Essay 5 to be equal across the two groups. We subtract the -2loglikelihood yielded by the restricted model from that obtained from the unrestricted and, remembering that the difference between two $\chi^2$ statistics is also $\chi^2$, we test that difference for significance; the number of degrees of freedom of the statistical test is equal to the number of parameters restricted. If it is not significant we would conclude that the extra flexibility gained by allowing different parameters for the focal and reference groups is not required—there is no DIF. The results from these analyses are shown in table 3 below. It yields a difference of 136. This figure is referred to a $\chi^2$ with 10 degrees of freedom. We must conclude that there is significant DIF against men for this essay. The average amount of this DIF is about 1.2 points (out of a maximum possible of 45) in the observed score scale, using the operational test's item weights.

Analysis 3 - Standard observed score scoring

The mean scores for men and women on the entire test for the 2,000 individuals examined in these three analyses are shown in table 4 below. Men who chose essay 5 scored approximately 9.3 points higher than women who chose that essay. When, in
analysis 2, we down weight the contribution to total score that is made by essay 5 we find that the difference between the mean scores by sex increases to 10.4 points. When we adjust for DIF in analysis 1 the difference between the sexes increases to 10.5 points. Thus we see that most of the increase in the precision of the proficiency estimate is obtained by weighting the essay commensurate with its informational content; the DIF adjustment gets us very little more.

----------------------------------------------------

Insert Table 4 about here

----------------------------------------------------

Discussion and conclusions

First, we must emphasize that we are not recommending that DIF be ameliorated through the use of scores estimated separately by group. This is illegal. We only calculated what such scores would be as a measure of comparison. Our suggestion is to consider reducing the DIF of an item by down-weighting its contribution to the total score. We used IRT weights as a natural way to accomplish this. Note that when each component of the test is IRT weighted, that is each item's contribution to total score is related to that item's relationship to the best fitting underlying latent dimension, two things occur. First the male-female difference increases a bit, but also the effect of essay 5's DIF shrinks to 8% of its size under standard weighting.

The reason for the larger difference between the sexes with IRT weighting is the relatively larger importance given to the MC section. The items in this section are more highly related to \( \theta \) than in either of the essay sections and so they are counted more. Since men do relatively better on the MC section than women this increases (slightly) the difference between the sexes. But, since essay 5 is not being counted quite so heavily, the DIF that it contributes is similarly down-weighted. Thus by using IRT weights we in-
crease the difference in scores between men and women, but decrease the DIF on the test. In addition the reliability of the test has increased from .83 to .93. This increase in reliability corresponds to what we would expect to occur if the test’s length more than doubled (actually Spearman-Brown would predict this if the test length increased by a factor of 2.6). Thus the test, scored in this way is simultaneously fairer and more accurate.

How can we say that a test that increases the differences between men and women is fairer? This is only true if the scores accurately represent differences in the knowledge of American history of the men and women who took this test. We can only fully resolve this question with carefully designed and executed validity studies. But an analogy at this point is helpful. Suppose we measure men's and women's height accurately and discover that there is a 6 inch difference on average. If we replaced the careful measurement with a random number we would find that there is no difference between the sexes. But finding 'no difference' is biased whereas finding 6 inches difference is correct and fair. Scores on Essay 5 are not random numbers, but they are analogous to using a rubber ruler that we must stretch a bit prior to each use. This adds a random component to all of the measurements and so reduces the measured difference between the sexes. We see the effect of this random component in the reduction of the test's overall reliability when Essay 5 is over-weighed with the standard weight. We would reach a different conclusion if the test is multidimensional and the essays are testing some characteristics of proficiency that are orthogonal to what is being tested by the MC items. There is little evidence on this exam, or other AP exams that have been examined, for multidimensionality to be much of a problem, but more light can be shed on this less likely possibility through validity studies.

There are some other possible solutions. In addition to the IRT path we have described, it is worthwhile to mention another road toward the same destination. If we feel wedded to the use of a priori weights associated with each item we can still find a way of devaluing the contribution of a somewhat flawed item. This can be done by reducing the
variability of scores (shrinking inward) on each item proportional to the value of that item. To illustrate this with an extreme example, suppose we have an item that was entirely worthless, that is suppose examinees were given scores between 1 and 10 on an entirely random basis\textsuperscript{7}, and in addition this item was to count 25\% of the final score. What can we do? Suppose we shrink all scores inward to, say five. We then add a five to everyone's score (weighted by 25\%). The item still counts 25\% but it exerts no ill effect on anyone's score. Empirical Bayes procedures provide one way to calculate the amount of shrinkage; such methods could be amended to include, as part of the error term, contributions due to DIF. Although we feel that such an approach has much to recommend it, we will leave a fuller exploration of it to a subsequent account.

A second solution, although not as elegant, is to maintain the current hybrid character of the tests, after all 1.2 points doesn't seem like very much\textsuperscript{8}. But be sure not to include too many such large items. So long as the test has a firm base of tried (that is pretested) and true items, a little noise and bias may be considered tolerable in return for the construct validity that it conveys. It is important however not to confuse such items with good measurement and to thence overpopulate the test with them.

The movement to include more large items in tests has, as its goal the increase of the test's validity. The IRT weighting scheme we have proposed maximizes the correlation of each item with a single underlying latent trait. That trait is abstracted from the particular combination of items included in the test. As such it will be heavily influenced by the more informative (Fisherean information per examinee minute) multiple choice items. This need not be the case. It would be a simple matter to choose item weights in such a way so as to maximize validity. But to do this we would need good validity-criterion data. Such data are not often available and so we used internal criteria to construct the item weights.
The proposal we have made works when the items that are found to have DIF are also the items whose connection to the underlying trait is more slender. Implicit in this is that there must be items with a relatively stronger connection to the underlying trait. It happens that the situation common in many contemporary tests fits this description. Multiple choice items are usually pretested and so those that make to an operational form do not typically display much DIF. Multiple choice items, lacking as they do, an error component associated with rater variation, tend to have a relatively strong relationship with the underlying trait. Essay questions, since they are not usually extensively pretested, have a stronger likelihood of displaying DIF. And, for reasons far broader then just the extra variation due to the subjectivity of scoring, they tend to have a weaker relationship with the underlying trait than the apriori weights usually assigned to them by test construction committees. This proposal is not likely to be of much help when a test is homogeneously constructed, say with all essays.

In summary, we believe that a test's performance can often be improved in both its precision and its fairness through the use of IRT weighting. However to accomplish these laudable ends requires abandoning a priori weights. Paraphrasing Jimmie Savage's famous observation about the benefits of Bayesian methods, "If you want to eat the IRT omelet you must be willing to break the IRT egg."
References


Figure Captions

Figure 1. A schematic representation of the structure of analysis 1. Females are considered to have omitted essay 5* and males essay 5**. The rest of the test is used as an anchor.

Figure 2. Box and whisker plots of the distribution of estimated scores by sex from IRT. The central circle represents the median, the outside circles are the extremes, the box covers the middle 50% of the scores.
Figure 1
IRT Scored w/ Adjustment for DIF on Choice Essay

Score on Whole Test

200
150
100
50
0

Male
Female

Figure 2
Table 1. Mean scores on the optional essays

<table>
<thead>
<tr>
<th>Essay</th>
<th>Men</th>
<th>Women</th>
<th>Difference</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>6.2</td>
<td>6.5</td>
<td>-0.3</td>
</tr>
<tr>
<td>3</td>
<td>5.5</td>
<td>5.4</td>
<td>0.1</td>
</tr>
<tr>
<td>4</td>
<td>6.0</td>
<td>5.7</td>
<td>0.3</td>
</tr>
<tr>
<td>5</td>
<td>6.5</td>
<td>5.8</td>
<td>0.7</td>
</tr>
<tr>
<td>6</td>
<td>6.1</td>
<td>5.5</td>
<td>0.6</td>
</tr>
</tbody>
</table>
Table 2. Conversion points for equipercentile equating

<table>
<thead>
<tr>
<th>Examination Grade</th>
<th>Percent at Grade</th>
<th>Theta</th>
<th>Raw Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>5</td>
<td>10.3</td>
<td>1.31</td>
<td>117</td>
</tr>
<tr>
<td>4</td>
<td>27.4</td>
<td>.33</td>
<td>91</td>
</tr>
<tr>
<td>3</td>
<td>28.0</td>
<td>-.40</td>
<td>72</td>
</tr>
<tr>
<td>2</td>
<td>27.6</td>
<td>-1.47</td>
<td>47</td>
</tr>
<tr>
<td>1</td>
<td>6.8</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Table 3. Likelihood Ratio Fit Statistics for analyses 1 and 2.

<table>
<thead>
<tr>
<th>Model</th>
<th>-2 Loglikelihood</th>
<th>Degrees of Freedom</th>
</tr>
</thead>
<tbody>
<tr>
<td>2. Restricted (No DIF)</td>
<td>960,870</td>
<td>370</td>
</tr>
<tr>
<td>1. Unrestricted (DIF)</td>
<td>960,734</td>
<td>360</td>
</tr>
<tr>
<td>Difference</td>
<td>136</td>
<td>10</td>
</tr>
</tbody>
</table>
Table 4. Mean scores by sex obtained through each analysis method

<table>
<thead>
<tr>
<th>Analysis</th>
<th>Male</th>
<th>Female</th>
<th>Difference</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. DIF</td>
<td>87.1</td>
<td>76.6</td>
<td>10.5</td>
</tr>
<tr>
<td>2. No DIF</td>
<td>87.1</td>
<td>76.7</td>
<td>10.4</td>
</tr>
<tr>
<td>3. Observed Score</td>
<td>86.9</td>
<td>77.6</td>
<td>9.3</td>
</tr>
</tbody>
</table>
Footnotes

§ This research was partially supported by the Advanced Placement Program of the College Board and we are pleased to be able to acknowledge their help. We are also grateful for some suggestions made on an earlier draft by Barbara Plake, David Thissen, Rebecca Zwick, and some wise but anonymous referees, although they should certainly not be held responsible for how we used them. Last, we would like to thank Claudia Gentile for her careful breakdown on the real cost of scoring an essay. This work was accomplished while the second author was an ETS Post-Doctoral Fellow; his current address is 5804 Zone 5, Pimville, Johannesburg 1808, Republic of South Africa.

1Barbara Plake has suggested that because the terminology for items as being "large" or "big" is non standard, we should define it more carefully. So here it is, by "large' or "big" we simply mean items, of whatever form, that because of their nature take so much examinee time that one can't have more than a few of them in any reasonable length test.

2It is often impractical (some have even suggested it is unethical) to insert a non operational test item into an operational test if that item takes up a major portion of the testing time. State Senator LaValle of New York has suggested that all 'experimental' (non operational) items must be so labeled on any test to allow the examinees to skip them if they wish. Such a practice is unlikely to yield unbiased estimates of the desired item characteristics. Extensive pretesting of large items is also often prohibitively expensive to score.

3A limited number of short items might be remembered from a pretest form and passed on to examinees who might have them on an operational test form. But limitations to memory, the use of multiple test forms, and the frequency with which experimental items do not qualify for operational forms, makes the effects of this exposure modest.
Large items (e.g. a specific essay prompt) are much more susceptible to the effects of prior exposure.

4 What does it cost to score an essay? At ETS we estimate that an essay that takes 30 minutes to write will be about 6 pages long. Good raters can score 10 such essays an hour and can keep up this pace for about 7 hours a day. The average cost of a rater (including salary, benefits, and overhead) is $170/day. This yields a cost for a single reading of an essay of $2.43. All essays are read at least twice, with additional readings required when there is a substantial difference of opinion between the two initial readers. A useful rule of thumb is that an average essay has 2.33 readings, yielding a cost for scoring a single essay of $5.67. This figure ignores training, setup time and various overhead expenses; including them boosts this rough estimate to $7 per essay. Tests with volumes over 100,000 easily accumulate six figure scoring expenses for a single essay. NAEP scores more than five million essays annually. Obviously real savings can be realized if we can use objectively scored items to test what we are interested in.

5 This corresponds to what Little and Rubin (1987) call 'missing at random;' an assumption that is almost certainly not true in this instance, but remains a pretty good place to start.

6 The reliability referred to here is generally called 'marginal reliability' obtained by integrating the error variance over the proficiency distribution. Details on how this is computed and why one might want to do it this way are contained in Sireci, Thissen, & Wainer (1991) and more thoroughly in Wainer & Thissen (1996).
Or worse suppose each examinee's score was chosen at random from one of two distributions with different means, and the distribution from which this random score was chosen was picked on the basis of sex or ethnicity.

We recognize that it would have been a more dramatic demonstration had we used an item with greater DIF. Happily for professionally prepared tests, such items are very difficult to find. We know, we looked. Thus our options were to fabricate a dramatic example, or use real data with a very modest effect. We opted for real, both because it provides a more accurate depiction of the size of the likely effect with carefully prepared tests, and to show the size of the effects that are detectable with model-based likelihood ratio procedures.

Although we could achieve much the same ends using a priori weights, if they were set from the results of IRT fitting of prior parallel AP test forms and not pulled out of someone's hat.