ED 466 712 TM 034 269

AUTHOR Milewski, Glenn B.; Baron, Patricia A.

TITLE Extending DIF Methods To Inform Aggregate Reports on Cognitive Skills.

PUB DATE 2002-04-00

NOTE 44p.; Paper presented at the Annual Meeting of the National Council on Measurement in Education (New Orleans, LA, April 2-4, 2002).

PUB TYPE Reports - Research (143) -- Speeches/Meeting Papers (150)

EDRS PRICE EDRS Price MF01/PC02 Plus Postage.

DESCRIPTORS *Analysis of Covariance; *Cognitive Ability; College Entrance Examinations; Higher Education; *Item Bias; Skill Development; *Test Items

IDENTIFIERS *Aggregation (Data); *Mantel Haenszel Procedure; National Merit Scholarship Qualifying Test; Preliminary Scholastic Aptitude Test; Standardization

ABSTRACT

Maximum usefulness of information provided to teachers and program staff by the enhanced score reports of the Preliminary SAT/National Merit Scholarship Qualifying Test (PSAT/NMSQT) depends on the development of aggregate skill reports; that is, reports that provide skill information aggregated across groups of students, such as grade level within high school. In this study, differential item functioning (DIF) methodology is extended to individual skill performance in order to compare aggregate groups, like schools or states, to the total population matched on an overall score. The results of four DIF detection methods used for this purpose are compared. Mantel-Haenszel and standardization procedures were calculated using binary skill classification data, and polytomous standardization and analysis of covariance (ANCOVA) were calculated using continuous posterior skill mastery probabilities. Results indicate that the ANCOVA approach was very sensitive to differential skill performance between the aggregate group and the population matched on ability. The Mantel-Haenszel, standardization, and polytomous standardization approaches provided consistent results that were less sensitive to differential skill performance. Taken together, these findings suggest that the ANCOVA approach may be too sensitive to differences in skill performance across groups matched on ability. Limitations and directions for future research are discussed. An appendix provides additional information about student skills. (Contains 10 tables and 22 references.) (Author/SLD)

Running head: AGGREGATE REPORTS ON COGNITIVE SKILLS

Extending DIF Methods to Inform Aggregate Reports on Cognitive Skills

Glenn B. Milewski

Fordham University

Patricia A. Baron

Educational Testing Service

Paper presented at the annual meeting of the National Council of Measurement in Education

in New Orleans, Louisiana, April 2002

Abstract

Maximum usefulness of information provided to teachers and program staff by the enhanced score reports on the PSAT/NMSQT® requires the development of aggregate skill reports; that is, reports that provide skill information aggregated across groups of students, such as grade level within high school. In the current study differential item functioning (DIF) methodology is extended to individual skill performance in order to compare aggregate groups, like schools or states, to the total population matched on an overall score. The results of four DIF detection methods used for this purpose are compared. Mantel-Haenszel and standardization procedures were calculated using binary skill classification data and polytomous standardization and the analysis of covariance (ANCOVA) were calculated using continuous posterior skill mastery probabilities. Results indicated that the ANCOVA approach was very sensitive to differential skill performance between the aggregate group and the population matched on ability. The Mantel-Haenszel, standardization, and polytomous standardization approaches provided consistent results that were less sensitive to differential skill performance. Taken together, these findings suggest that the ANCOVA approach may be too sensitive to differences in skill performance across groups matched on ability. Limitations and directions for future research are discussed.

Extending DIF Methods to Inform Aggregate Reports on Cognitive Skills

The College Board administers the Preliminary SAT®- National Merit Scholarship Qualifying Test, or PSAT/NMSQT, to over 2 million students each year. The PSAT/NMSQT measures verbal reasoning, math reasoning and writing skills. Students typically take the PSAT/NMSQT in their junior year to: a) prepare for other standardized exams (e.g. SAT I, SAT II, etc.), b) qualify for the National Merit Scholarship Corporation's (NMSC) scholarship programs, c) compare scores to other college bound students, d) forecast future SAT scores, e) assess verbal, math, and writing skills, f) participate in the Student Search Service to get information from colleges, and g) get feedback on academic strengths and weaknesses (www.collegeboard.org, 2001). In recent years The College Board's method of reporting achievement on the PSAT/NMSQT shifted focus to include a diagnostic profile of cognitive skills for every test taker in addition to a report of scaled scores and item performance.

Cognitive Skills

The PSAT/NMSQT diagnostic profile includes information about how well test takers perform on the underlying knowledge and cognitive processing skills, or attributes, required for answering questions correctly (DiBello, 2002). The PSAT/NMSQT diagnostic profile is produced through an application of a modified version of the rule-space model (Tatsuoka, 1983; Tatsuoka & Tatsuoka, 1992) developed by DiBello (2002)[1]. This modified rule-space model serves as a statistical method for classifying students' item responses into a set of attribute-mastery patterns associated with different cognitive skills (Gierl, Leighton, & Hunka, 2000).

[1] Since the modifications to the rule-space model applied to the PSAT/NMSQT diagnostic profile are quite extensive in certain instances, the reader should regard the current section of the paper as only a general description of the model providing a global definition of how cognitive skills are measured. For a more exact treatment of the diagnostic classification method applied to the PSAT/NMSQT, the reader is referred to DiBello (2002).

These attribute-mastery patterns are used to classify a test taker as having mastered or not mastered a particular knowledge state, or skill.

The data used in the current paper are of two types: posterior mastery probabilities and binary classifications. Because the current study is not investigating the theoretical work on which these data are based, the application of the modified rule-space method will be described here briefly. For a more thorough discussion of the modified rule-space, please see DiBello and Crone (2001) and DiBello (2002).

The first step in applying the rule-space methodology is to identify the cognitive attributes to be measured by a test. Once the attributes are identified, the pool of test items that are created can be mapped onto the attributes in order to produce what it generally referred to as the Q incidence matrix. The Q incidence matrix is of the form I × K, where I refers to the number of items and K refers to the number of attributes. The elements of the matrix are zero and one, where zero represents the absence of a particular item for an attribute and one represents the presence of a particular item for an attribute. Let $X = (X_1, \ldots X_{i)},$ denote the item response pattern, where $X_i = 0,1$ represent incorrect and correct response to item i; and let $\alpha = (\alpha_1, \ldots \alpha_{K)}$, denote the attribute pattern or knowledge state, where $\alpha_K = 0,1$ stand for non-mastery/mastery of attribute k. The posterior probability is the probability of a certain knowledge state given a pattern of item responses (i.e. the probability that $\alpha_K = 1$ given X).

The rule-space method is based on a conjunctive latent response model which posits that test takers who have mastered the cognitive attributes required for an item, found by reference to the Q matrix, will answer that item correctly. For any knowledge state α, an ideal response pattern Z^α (read as zeta super-alpha) can be associated with α, where $Z_i^\alpha = 1$ if all attributes required by item i are mastered in α, otherwise $Z_i^\alpha = 0$.

Define a map $P(X) = (\theta_x, Z_x)$; where θ_x is IRT theta and z_x is the standardized IRT caution index (Tatsuoka, Linn, Tatsuoka & Yamamoto, 1988; Tatsuoka, 1997). Define θ_α to be the θ associated with θZ^α, so $\theta_\alpha = \theta_Z\alpha$; $\zeta_\alpha = \zeta_Z\alpha$. The probability distribution $P(X|\alpha)$ is induced from a normal distribution in θ,Z space centered at θ_α , $\zeta_{\alpha.}$ That distributional assumption allows for calculation of the probability of a certain response vector X conditional on a knowledge state alpha.

Using as input the item response string (11101....1), posterior probabilities of mastery can be calculated for each skill P(alpha_k=1 | X). The classification of students as N, M or U occurs after posterior mastery probabilities have been calculated, where M = skills likely mastered, N = skills to improve (non-master) and U = unknown). Classification strings (NUNMMN,...N) will be converted into binary data, by mapping N = 0 and both M and U = 1; determination of which skills should be classified as needing improvement was the goal in the dichotomization.

The posterior probability is an indicator for mastery on a cognitive attribute. The values for posterior probabilities on cognitive attributes are important for informing researchers on the strength of evidence that test takers have mastered the cognitive skills measured by a test. These probabilities can be translated into a simple binary classification, mastery or non-mastery. Typically, if a test taker's probability on a given cognitive attribute falls below a cut point on the distribution of probabilities for an attribute and, if he or she has responded incorrectly to more that one item associated with that attribute, then the test taker is considered to not have mastered that particular knowledge state. Test takers with all other posterior probabilities are considered to have either mastered the cognitive attribute or to have an undetermined mastery status.

In October of 2001 the modified rule-space model (Crone & DiBello, 2001; DiBello, 2002) was applied to the item responses provided by PSAT/NMSQT examinees. The results of this application were summarized in enhanced score reports provided to students. The enhanced score report contained a diagnostic profile of each students' cognitive skill performance, in addition to a summary of test, sub-scale and item performance. In order to maximize the usefulness of the enhanced score reports provided to students, it is necessary to provide teachers and administrators with an aggregate report that summarizes their students' performance on skills.

A general statistical framework for aggregate reporting involves extending differential item functioning (DIF) methods to compare the performance of students within an aggregate group to the performance of the population of students matched on ability. In the current study, this framework was applied to cognitive skills rather than to test items, so it may be more appropriate to use the term differential *skill* functioning (DSF) rather than DIF. Since the statistical methods remain the same regardless of whether they are applied to item responses or skill performance the term DIF will be used here, but explanations of DIF methods will be couched within the context of aggregate reporting on cognitive skills. Additionally, there is an important distinction in the interpretation of DSF versus DIF.

Traditionally, DIF methods are used to examine item response differences across a focal group comprised of a subgroup of examinees (e.g. a group of ethnic minority students) and a reference group comprised of the focal group's counterpart (e.g. White students). The common interpretation of DIF indicators is to identify potential bias in items, such that the item favors either the focal group or reference group and is measuring something irrelevant to the construct of the test. When such items are identified, they are often excluded from a test.

When DIF methods are extended to aggregate reports on cognitive skills, the aggregate group of interest, like students within a grade level at a particular high school, becomes the focal group and a random sample of the population becomes the reference group. After these groups are matched on overall ability, DIF methods function to reveal which skills the aggregate group performed the same, better, or worse than the population of examinees. Under this use, significant differences in skill performance reveal that the aggregate group (e.g. juniors in a high school) did not perform as expected on a skill given that the aggregate group and reference group are matched on ability. These results provide the aggregate report user with information about skill performance in the aggregate group, and not information regarding whether a skill is biased.

The current study focused on comparing four DIF detection methods used for aggregate reporting on cognitive skill performance. Since the rule-space model produced continuous posterior skill mastery probabilities and binary skill classifications, both binary and polytomous DIF detection methods were compared. Two DIF detection methods for aggregate reporting on binary skill classifications, Mantel-Haenszel and standardization, were calculated. Two DIF detection methods for aggregate reporting on continuous posterior skill mastery probabilities, polytomous standardization and the analysis of covariance (ANCOVA), were also calculated. The next subsection provides a more detailed description of each DIF detection method examined in the current study.

DIF Statistics

For each of four statistical methods employed in the current study, a conceptual and mathematical definitions as well as explanations of effect size indices are provided. Further, interpretations of the statistical criteria for flagging whether a DIF effect size index indicated that an aggregate group performed significantly better, worse, or the same on a skill as the reference group matched on ability is also explained.

The Mantel-Haenszel Approach

The Mantel and Haenszel (1959) approach evaluated the null hypothesis that there was no difference in the odds of mastering a skill between the aggregate group and the population of examinees. The null hypothesis was rejected when the odds of mastering a particular skill in an aggregate group were statistically different from the odds of mastering that skill in the population. For example, the null hypothesis was rejected if the odds of mastery for a skill were significantly better, or significantly worse, in the aggregate group then they were in the population at every level of matching variable. The matching variable used in the current study was the total score on the Math section of the PSAT/NMSQT®.

The null hypothesis associated with the Mantel-Haenszel approach was evaluated with an effect size known as Mantel-Haenszel Delta DIF (MH D-DIF). Computationally, MH D-DIF involves finding the ratio of the odds of skill mastery between the aggregate and population, taking the log of this ratio, and scaling the resulting value so that it is on a skill difficulty metric. The term delta is used because it refers to an item difficulty metric developed by Educational Testing Service (ETS) that has a mean of 13 and a standard deviation of four.

In order to fully describe the Mantel-Haenszel approach employed in the current study, it must be defined mathematically. MH D-DIF uses the Mantel and Haenszel (1959) estimate of the odds ratio (α_{MH}), which is defined as:

(1) $$\alpha_{MH} = \left[\sum_m R_{rm} W_{fm} / N_{tm}\right] / \left[\sum_m R_{fm} W_{rm} / N_{tm}\right]$$

where, R_{rm} is the number of examinees with mastery status in the population at m.

W_{fm} is the number of examinees with non-mastery status in the aggregate group at *m*.

N_{tm} is the total number of examinees in both the aggregate group and the population at *m*.

R_{fm} is the number of examinees with mastery status in the aggregate group at *m*.

W_{rm} is the number of examinees with non-mastery status in the population at *m*.

Once the odds ratio is calculated, MH D-DIF can be calculated using the following formula provided by Dorans and Holland (1993):

$$MH\ D\text{-}DIF = -2.35\ ln[\alpha_{MH}]. \quad (2)$$

For interpretative purposes, negative MH D-DIF values indicated that a skill was more difficult in the aggregate group whereas positive MH D-DIF values indicated that a skill was easier in the aggregate group. Large negative MH D-DIF values showed that skill performance was worse in the aggregate group and large positive MH D-DIF values showed that performance was better in the aggregate group.

Dorans and Holland (1993) and Zieky (1993) developed statistical criteria for flagging MH D-DIF values that constitute a large difference across groups. These criteria, while originally developed for identifying problematic items, are equally applicable for identifying skills in which aggregate groups have performed significantly better or worse. These criteria were adapted for use in aggregate skill reporting so that skill performance between the reference and focal group was the same (a) (e.g. differences were negligible) when MH D-DIF had an absolute value less than 1.0, moderately different (b) when MH D-DIF had an absolute value that was greater than 1.0, largely different (c) when MH D-DIF had an absolute value that was greater than 1.5. These statistical criteria for flagging skill provide a means of organizing how skill performance is reported to aggregate groups.

The Standardization Approach

The standardization approach examined whether the expected mastery on a skill differed for examinees of equal ability from different groups. When skill performance was measured through binary classification, expected mastery corresponded to the proportion at mastery in the aggregate group and the population. Expected mastery can be further operationalized as the nonparametric skill test regression, where differences in empirical skill test regressions are indicative of the aggregate group performing better or worse on a skill than the population matched on ability (Dorans & Holland, 1993). The null hypothesis associated with the standardization states that the empirical skill test regressions are equivalent across the aggregate group and the population:

(3) $$H_o\!: E_f(S\,/M) = E_r(S\,/M)$$

where, $E_f(S\,/M)$ is the empirical skill test regression for the aggregate group

$E_r(S\,/M)$ is the empirical skill test regression for the population

S is the skill mastery variable

M is the matching variable.

The standardization approach enables skill test regressions to be plotted and differences in skill performance across groups matched on ability to be analyzed visually (see Dorans & Holland, 1993). A visual analysis of this kind of plot would lead to a rejection of the null hypothesis if skill test regressions were different across groups.

More commonly, the null hypothesis associated with the standardization approach is evaluated with an effect size known as the standardized p-difference (STD P-DIF). Dorans and Holland (1993) defined STD P-DIF mathematically with following equation:

(4)
$$STD\ P\text{-}DIF = \sum_m N_{fm}(P_{fm} - P_{rm}) \Big/ \sum_m N_{fm}$$

where, N_{fm} the number of examinees at m in the aggregate group

P_{fm} the proportion of mastery at m in the aggregate group

P_{rm} the proportion of mastery at m in the population group.

The formula above shows that computationally STD P-DIF is the weighted difference in expected mastery between the aggregate group and the population at every level of the matching variable. The differences in mastery are weighted by the number of examinees in the aggregate group at every level of the matching variable. Of particular importance to STD P-DIF is that the greatest weight is given to differences in P_{fm} and P_{rm} at those score levels most frequently attained by the aggregate group under study (Dorans & Holland, 1993). This quality functions to provide a robust estimate differences in binary skill classifications across groups.

In the current study, STD P-DIF values described DIF in terms of the difference in the proportion at mastery between the aggregate group and the population matched on ability. Positive STD P-DIF values indicated that performance on a skill was better in the aggregate group whereas negative STD P-DIF values indicated that performance on a skill was worse in the aggregate group. According to Dorans and Holland (1993) STD P-DIF values that are outside the range of $\{-0.10,+0.10\}$ are more unusual and should be examined very carefully. The current study adapted this criteria so that a skill was flagged as significantly better in the aggregate group if STD P-DIF was greater than 0.10. Conversely, a skill was flagged as significantly worse in the aggregate group if STD P-DIF was less than -0.10. All other STD P-

DIF values indicated that performance on a skill was the same between the aggregate group and the population matched on ability.

The Polytomous Standardization Approach

Dorans and Schmitt (1991) expanded the binary standardization approach so that the method could be applied to polytomous items. This expansion resulted in the development of the polytomous standardization approach, a method particularly useful for aggregate reporting on continuous posterior skill mastery probabilities.

When applied to aggregate reporting on cognitive skills, the polytomous standardization approach is essentially the same as the binary standardization approach. Again, the null hypothesis of no difference between the empirical skill test regressions across the aggregate group and the population is evaluated. The polytomous standardization approach diverges from the binary standardization approach only in that expected mastery corresponds to the mean on the posterior probability of skill mastery in the aggregate group and the population.

In the current study the effect size used for polytomous standardization was labeled as the standardized mean difference (SMD). The SMD effect size was defined mathematically as:

(5) $$SMD = \sum_m N_{fm}(E_{fm} - E_{rm}) \Big/ \sum_m N_{fm}$$

where, N_{fm} the number of examinees at *m* in the aggregate group

E_{fm} the posterior probability mean at *m* in the aggregate group

E_{rm} the posterior probability mean at *m* in the population.

This formula for the SMD computed the weighted difference in expected mastery between the aggregate group and the population at every level of the matching variable.

The SMD effect size describes DIF as the difference in the mean of the posterior probability of mastery between the aggregate group and the population matched on ability. A major advantage of SMD is that it provides a conceptually simple DIF detection index that is expressed in a metric consistent with the scale of the continuous variable (Penfield & Lam, 2000). Since the observed range of posterior probabilities is significantly smaller than the full range (0,1), the flagging criteria employed in the current study for SMD was set at 0.05. Performance was labeled as better in the aggregate group if SMD values were greater than 0.05, worse if SMD values were less than –0.05, and the same as the population for all other SMD values.

The Analysis of Covariance (ANCOVA) Approach

ANCOVA combines regression analysis with analysis of variance so that differences in a dependent variable across levels of a fixed factor can be estimated after controlling for a source of common variation, or covariate (Kirk, 1995). The ANCOVA approach can address the same purposes as other DIF detection methods because it captures differences in matched groups. When considered in the context of aggregate skill reporting, the analysis of covariance (ANCOVA) tests the null hypothesis that there is no difference in cognitive skill performance between the aggregate group and the population after taking into account differences in overall ability across groups. This null hypothesis can be tested with a variety of criteria including an overall F-test, an effect size, or a confidence interval. In order for these criteria to be unbiased, several statistical assumptions must be met.

ANCOVA requires that a set of five assumptions be met including, (a) independence of observations, (b) normality on the dependent variable, (c) homogeneity of variance in the dependent variable across level of the fixed factor, (d) linearity between the dependent variable and the covariate for all levels of the fixed factor, and (e) parallelism of the regression of the covariate on the dependent variables across all levels of the fixed factor (see Keppel, 1991 for a more complete review). In order to more fully meet the assumptions of ANCOVA, particularly assumptions (b), (d), and (e), it is sometimes necessary to transform the scale of the dependent variable.

Frequently, probability distributions are non-linear because there are fewer cases at the tails of the distribution. The logit transformation $\{\ln(p/(1-p)\}$ addresses this non-linearity by stretching the probability distribution at the tails. Since the current study used the ANCOVA approach to address matched differences in non-linear continuous posterior skill mastery probabilities, it was necessary to apply the logit transformation in order to more fully meet the linearity and parallelism assumptions.

In the current study, the dependent variable used for the ANCOVA approach was the continuous posterior skill mastery probability, the fixed factor was the grouping variable with two levels- the aggregate group and the population, and the covariate was the total score on the associated sub-scale of the skill considered in the analysis. The general formula for ANCOVA used in the current study, as it applied to aggregate reporting on cognitive skills, was:

(6) $$\hat{Y}_{ij} = \hat{\beta}_T\left(X_{ij} - \bar{X}_{..}\right) + \bar{Y}_{..}$$

where, $\hat{Y}_{ij}$ is the predicted score on the continuous posterior probability of skill mastery

$\hat{\beta}_T$ is the linear regression coefficient computed for the $i = 1, \ldots, n$ and $j = 1, \ldots, p$ pairs of X and Y scores

X_{ij} is the covariate, or sub-scale score, for subject i in treatment level j

$\bar{X}_{..}$ is the mean of the covariate

$\bar{Y}_{..}$ is the mean of the continuous posterior probability of skill mastery.

This formula illustrates how regression analysis and the analysis of covariance are combined to provide a statistically powerful method for aggregate reporting on cognitive skills.

The results of ANCOVA were used to describe whether skill performance in the aggregate group was better, worse, or the same as performance in the population after accounting for overall ability. The effect size measure used to describe differences in continuous posterior mastery probabilities between the aggregate group and the population matched on ability was the unstandardized *b* coefficient. The unstandardized *b* coefficient captured the effect of the difference between groups after accounting for the influence of the covariate. Since the logit transformation was applied, values on the unstandardized *b* coefficient were interpreted as the log odds ratio ($\log(\hat{\theta})$) between the aggregate group and the population matched on ability (C. Lewis, 2001, personal communication). Interestingly, since the unstandardized *b* coefficient is equivalent to a log odds ratio it can be placed on the delta scale if it is multiplied by –2.35.

The 95% confidence interval for the unstandardized *b* coefficient was also calculated so that the value of unstandardized *b* coefficient was not the sole description of ANCOVA DIF in the current study. The 95% confidence interval for ($\log(\hat{\theta})$) was found by adding and subtracting two times the standard error of the ($\log(\hat{\theta})$) estimate. Agresti (1996) provided a formula for the asymptotic standard error of the log odds ratio ($ASE(\log(\hat{\theta}))$), which entailed taking the sum of the inverse of cell counts in the focal and reference group at every level of *m* to the half power:

(7) $$(ASE(\log(\hat{\theta}))) = \left(\sum_m \frac{1}{n_{fm}} + \frac{1}{n_{rm}}\right)^{1/2}$$

where, $\frac{1}{n_{fm}}$ is the inverse of the number of cases in the focal group at *m*

$\frac{1}{n_{rm}}$ is the inverse of the number of cases in the reference group at *m*.

The confidence interval around ($\log(\hat{\theta})$) shrinks, or becomes more precise as the number of cases across groups increases.

In the current study, the unstandardized *b* was interpreted as the log odds ratio between the aggregate group and the population. Negative values on the raw unstandardized *b* coefficient indicated that skill performance was better in the aggregate group and positive values indicated that performance was worse in the focal group. This pattern of interpretation of positive and negative values on the unstandardized *b* coefficient was the exact opposite of the other DIF detection effect sizes described so far (MH D-DIF, STD P-DIF, and SMD) where positive values showed better performance in the aggregate group and negative values showed worse performance in the aggregate group.

The following criteria were used to flag whether skill performance was significantly better or worse in the aggregate group. First, the raw unstandardized *b* coefficient and its 95% confidence interval was transformed through multiplication by –2.35 so that the values were on the delta scale. A value of one on ($\log(\hat{\theta})$) on the delta scale is generally considered to show no difference between groups. Second, if a transformed confidence interval did not contain one *or* an absolute value on a transformed unstandardized *b* was less than one, skill performance in the aggregate group was considered the same (a) as skill performance in the population. Third, if the transformed unstandardized *b* produced a confidence interval that did not contain one *and* its absolute value was greater than 1.5, skill performance was considered to be largely different (c) between the aggregate group and the population matched on ability. Fourth, all other transformed values were considered to exhibit moderate (c) differences in skill performance across the aggregate group and population matched on ability.

DIF and Aggregate Reports on Cognitive Skills

The applicability of DIF methods to the measurement of cognitive skills lies in the potential for the results to inform aggregate reports on cognitive skills. The current section of the paper defines aggregate reports and outlines steps for using DIF to inform the results provided in an aggregate report.

An aggregate generally represents a group of persons that have certain traits or characteristics in common without necessarily having any direct social connection with one another (Vogt, 1999). Some examples of aggregate groups include schools, districts, and states. An aggregate report provides information not at the individual level, but at the aggregate, or group level. Diagnostic profiles of cognitive skills for PSAT/NMSQT examinees create a need for summary reports of the information provided by these profiles at grade level within the schools, districts and states level.

Differential item functioning (DIF) detection procedures yield statistical information that is useful for constructing this kind of aggregate report. If aggregate groups are compared to the national sample, DIF procedures can be applied using scaled scores, to represent overall ability, as the matching criterion. This allows for comparisons of performance on items or skills between small samples, such as juniors within a school, to a larger matched group so that items or skills on which performance is above or below expectations can be identified. There are several steps to apply when using DIF detection methods for aggregate reporting on cognitive skills measured by the PSAT/NMSQT.

In the case of the current study using PSAT/NMSQT data, the level of aggregation will be grade within school and grade within state. This decision is based on the needs of the users of PSAT/NMSQT scores. The reference group is based on the national group of sophomores and juniors who took one form or version of the test, a population of over 1 million students. For illustrative purposes, let's say the researcher chose 11th grade examinees from Acorn High School, Any town, USA as the aggregate group. The researcher could use DIF methods to compare 11th grade student performance on Math in this fictitious high school to a reference group matched on ability. In the current example, the reference group would be comprised of a sample of all 10th and 11th grade examinees from the test administration matched on the scaled score for math. It is important to note that in order to ensure accuracy in the results of the DIF detection method, no less than 25 students should comprise the focal group. Preferably, however, the focal group should be comprised of at least 50 examinees (Zieky, 1993).

In applying DIF for aggregate reporting on cognitive skills the goal is to report to the aggregate group the skills in which their group has performed significantly worse, or better, than the reference group matched on ability. First, the estimate of DIF must be calculated on either the binary skill classification or the continuous posterior skill mastery probability for all skills in a common content area (e.g. math, verbal, or writing). Then, the values must be examined to determine whether differences are large enough to be of concern. In the case where an aggregate group is performing better than or worse than one would expect, given the performance of a matched group, some flagging is necessary. Skills that are not flagged show that performance on a skill is the same in the local group. Skills that are flagged reveal that performance on a skill is better or worse in the aggregate. For example, positive values on MH D-DIF, STD P-DIF, and SMD that meet the criteria for flagging indicate that skill performance is better in the aggregate group whereas negative and flagged values on the unstandardized *b*, applied to logit-transformed probabilities, show that performance on a skill is better in the local group.

Once it is determined that for some skills the performance of the aggregate group differs from what might be expected, a graphical display may prove useful. Figure 1 provides an example of what a graphical aggregate report of performance in Acorn High School might look like. The scale of the y-axis is on the proportion mastery metric, which aids interpretation. The shaded portion of Figure 1 is an average error band found by taking the average standard error of a set of STD P-DIF estimates, multiplying it by two, and adding and subtracting the resulting value to zero. Skills that are plotted above the shaded portion indicate that skill performance is significantly better whereas skills plotted below the shaded portion show that skill performance is significantly worse in the aggregate group compared to the reference group matched on ability. Figure 1 represents a prototype of one possible way of reporting aggregate skill performance graphically. It is based on method for reporting aggregate PSAT/NMSQT item performance employed by ETS (see Lawrence, 1989[a]; Lawrence, 1989[b]; the College Board, 1990). Note that what is plotted in Figure 1 is item performance, not skill performance, hence the letters A, S and R representing item types. Other graphical methods, like box and whisker plots of DIF and error estimates for each skill may provide a more precise representation of differences in skill performance, but may present interpretation issues.

In addition to summarizing skill performance, the aggregate report should suggest methods for improving examinee performance on the skills for which the aggregate group performed significantly worse than the reference group matched on ability. In this way, the results of aggregate reports on skill performance become more fully linked to educational instruction.

The current study focused on comparing the utility of four DIF detection procedures that can be used to inform aggregate reports on cognitive skills measured by the math sections of the PSAT/NMSQT. The four DIF detection procedures were compared in eight high schools and one state that differed with respect to average ability, variability around average ability, and number of examinees tested.

Method

Participants

From the population of approximately 1.5 million sophomore and junior non-handicapped PSAT/NMSQT Form 1 examinees, 10 separate samples were selected. Each sample was deliberately selected to meet certain criteria. For example, a large random sample of examinees was selected to represent the reference group. In addition, nine focal groups were selected. The nine focal groups included one group comprised of examinees from one state (Focal group A) and eight groups comprised of examinees from eight different high schools (Focal groups B-I). Focal groups B through I were chosen so that comparisons in DIF detection methods could be made across sample size (small-large), average Math score (low-high), and variability in Math scores (low-high). Focal group A was chosen so that DIF methods for aggregate reporting could be investigated in a large state system. A complete listing of descriptive statistics for these groups of examinees is provided (see Table 1). For comparative purposes, the mean and standard deviation of the population was 46.5 and 10.8, respectively, in a scaled scored distribution that ranged from 20 to 80. All of the examinees from this sample received a valid Math Reasoning score on the exam.

Materials

The PSAT/NMSQT contains five sections: two 25-minute verbal sections, two 25-minute math sections, and one 30-minute writing skills section (www.collegeboard.com, 2002). Only the skills measured by the math sections were investigated in the current study. Forty questions comprise the two math sections. The formats of these questions are multiple choice, quantitative comparison, and student produced response. Responses to the questions from the October, 2001 PSAT/NMSQT administration were used to estimate performance on a total of 16 cognitive skills related to mathematics. Math skill number 12 was not reported due to the fact that no skills with fewer than 3 items are reported on any form. A complete listing of these skills is available (See Appendix A).

Procedure

Four DIF detection methods were applied to the 15 math skills measured by the October 2001 PSAT/NMSQT in each of the nine schools considered in the current study. The MH D-DIF and STD P-DIF statistics were calculated on the 15 math binary skill classifications for the nine schools in the current study. SMD and ANCOVA statistics were calculated on the 15 posterior skill mastery probabilities for the nine schools in the current study.

Results

The current study compared the results of four DIF detection methods within and across nine focal groups. The groups differed with respect to three factors (a) sample size, (b) average Math score, and (c) variability in Math scores. A complete listing of these results is provided in Tables 2 through 10. It is important to note that the raw unstandardized *b* coefficients and their 95% confidence intervals are reported in tables 2 through 10 and not the coefficients transformed by multiplication by –2.35. In addition, since the scale of the DIF effect sizes were different across methods, effect size values were categorized using the flagging criteria described previously. The following sub-sections describe the major findings for each of the three factors listed above.

Sample Size

Focal groups B, D, E and G had small samples. An exploration of the results for these groups yielded several important findings. When variability around the average Math score was low, ANCOVA procedure flagged more skills for DIF; when variability around the average Math score was high, MH D-DIF and STD P-DIF flagged more skills for DIF. One problem with analyzing the results was due to the fact that MH D-DIF could not be calculated when there was no variation in binary skill classification (i.e. when all members of a focal group mastered a particular skill). A lack of any variation in binary skill classification tended to occur when focal group samples were small, especially when variability around the average test scores in these groups was low.

Focal groups A, C, F, H, and I had large samples. Overall, the ANCOVA procedure flagged more skills for DIF, and values of SMD were very small for all groups with large N. The standardization approach, regardless of whether it was applied to binary skill classifications or continuous posterior probabilities, did not result in very many skills being flagged for DIF. Therefore, the results of ANCOVA and standardization were very often inconsistent. Similarly, STD P-DIF values were not flagged for any skills in the large N groups.

Average Math Score

Focal groups A, E, G, H, and I had high average math scores. For these groups, when sample size was low, inconsistencies were found in which skills were flagged for DIF depending on whether binary skill classifications or continuous posterior probabilities were evaluated.

Focal groups B, C, D, and F had low average math scores. Results indicated that when variability was high, STD P-DIF flagged more skills for DIF. On the other hand, when variability was low, the ANCOVA procedure flagged more skills for DIF.

Variability in Average Math Score

Focal groups B, F, G, and H had high variability in average math score. Results indicated that when samples were large, ANCOVA detected DIF in more skills. When samples were small, DIF detection methods applied to binary skill classification flagged slightly more skills for DIF.

Focal groups A, C, D, E, and I had low variability in average math scores. For these groups, across differences sample size, ANCOVA flagged most skills for DIF. Mantel-Haenszel could not be calculated in groups where all examinees either mastered, or did not master, a particular skill.

Discussion

The following results summarize the major findings of the current study. First, the ANCOVA approach, applied to continuous posterior probabilities of skill mastery, was generally more sensitive to differences in skill performance than the other DIF detection methods considered in the current study. Second, the sensitivity of the ANCOVA approach resulted in more skills being flagged for significantly better or worse skill performance in the aggregate group than MH D-DIF, STD P-DIF, and SMD. Third, since the results of MH D-DIF, STD P-DIF, and SMD described differences in skill performance consistently across aggregate groups, the results of ANCOVA were inconsistent with the other methods applied to the same data. In particular, ANCOVA results appear to be ascribing more differences in skill performance when total score variability was low.

Taken together, the major findings suggest that the ANCOVA approach may be too sensitive to accurately detect differences in skill performance between the aggregate group and the population matched on ability. It may also be the case that the ANCOVA approach produced inaccurate results. It is extremely important that the statistical methods used in aggregate reports on cognitive skills are accurate, since these reports may be influential for informing policy changes on funding for educational initiatives, changing classroom teaching methods, and a host of other practices for improving student success. Therefore, we do not recommend using the ANCOVA procedure for aggregate reporting on cognitive skills at this time.

Several reasons might account for the possible inaccuracy of the ANCOVA approach. For example, the results may have been inaccurate because of the criteria used to flag a skill as significantly better or worse in the aggregate group compared to the population matched on ability. It may be that other cut-off criteria for the unstandardized *b* and its 95% confidence interval should have been used for flagging skills. For example, the 99% confidence interval may be more appropriate.

On the other hand, the possible inaccuracy of the ANCOVA approach may have been due to the data not meeting the assumptions for the approach adequately. However, the parallelism and linearity assumptions, which are most important, were adequately met in nearly all of the aggregate groups once the logit transformation was applied to the posterior probability distribution. Only in certain situations, such as when aggregate groups were comprised of very small samples, were these assumptions not met as precisely. One solution to this potential problem involves fitting the ANCOVA model with multiple covariates that are polynomials of the Math scaled score variable (C. Lewis, 2002, personal communication). In order to avoid multicollinearity, these polynomials should be orthogonal. Still yet another reason for why the

ANCOVA approach may have been inaccurate concerns the nature of the skill performance data. The ANCOVA approach may be ill equipped to deal with the idiosyncrasies of the posterior probabilities distribution.

These same idiosyncrasies in the posterior probability distribution may account for why the SMD approach was less sensitive to flagging skills for significantly different performance in the aggregate group. On the other hand, it may be the case that the SMD cut-off criteria for flagging a skill for significantly different performance, {-0.05; 0.05} was too stringent. Either way, there are obvious problems with the current implementation of the SMD approach and so it too is not recommended for use in aggregate reports on cognitive skills. This recommendation, combined with the recommendation not to use the ANCOVA approach in aggregate reports, is perplexing to the current authors. Originally our intuitions were that using the continuous posterior mastery probabilities of skill mastery, over the binary skill classifications, would provide a more precise estimate of group differences. The results seem to indicate the opposite pattern.

With respect to the findings from the methods applied to binary skill classifications, the following conclusions are made. First, there are problems with both the Mantel-Haenszel approach and the standardization approach. For example, the odds ratio for MH D-DIF cannot be calculated when all members of an aggregate group master, or do not master, a particular skill. This kind of situation is not uncommon in aggregate reporting of PSAT/NMSQT results. Frequently aggregate groups are comprised of less than 50 students. The current implementation of the standardization approach to binary skill classifications was problematic as well since it did not employ a smoothing procedure and as a result, may have produced STD P-DIF values that were influenced by sampling error. A standardization approach that uses kernel smoothing of the empirical item response functions (see Lyu, Dorans, & Ramsay, 1995), such as the approach available within the TestGraf system (Ramsay, 1995), may offer an improvement to the potential influence of sampling error. Of the two approaches applied to binary skill classifications, STD P-DIF seems to be preferred method at this time.

As is the case in most research, the recommendations provided so far are tentative. The results of the current study, as well as the limitations, and other independent research ideas, create a need for future work to more fully address the relative advantages and disadvantages of applying DIF detection methods to generate cognitive skill reports at the aggregate level. An especially important area for additional research, involves exploring how the nature of the posterior probability distribution might influence DIF detection methods applied for the purposes of aggregate reporting.

Limitations and Future Research

First, a lack of analyses on simulated data, where the 'true' differences between groups was established *a priori*, limits the generalizability of the results and conclusions reached in the current study. Since we did not know the true difference between groups, we were unable to evaluate which DIF detection methods captured group differences most accurately. Instead, we could only evaluate consistency across DIF detection methods. Our next research step will be to perform a simulation study.

Second, the only characteristics that varied across aggregate groups were sample size, total or scaled score, and variability around total score. There is undoubtedly other important group characteristics that might influence the accuracy of DIF detection methods applied to cognitive skills. Some examples of these characteristics include (a) the distribution of performance on skills in the focal and reference group, and (b) the ratio of focal to reference group members, among others. In future research on the use of DIF for aggregate reporting on cognitive skills, these and other characteristics should be explored.

Third, the current study focused only on math skills so the results may not be generalizable to the Verbal and Writing skills measured by the PSAT/NMSQT. Future research should explore the relative performance of DIF detection methods applied to aggregate reports on Verbal and Writing skills. It may be the case that differences in the way performance on Verbal and Writing skills is estimated influences the accuracy or sensitivity of the DIF detection methods.

Fourth, inconsistencies in the direction of DIF (i.e. whether the focal or the reference group performed better on the skill) were found depending on whether binary skill classifications or continuous posterior probabilities were evaluated. Future research should be directed toward understanding why these inconsistencies were found.

Fifth, the methods for aggregate reporting on cognitive skills outlined in the current study are only appropriate when focal groups have a sample that is greater than 25 to 50 cases; in fact, further research using simulations may determine these sample sizes too small. An important research question is how can performance on cognitive skills be summarized more accurately for groups that have 25 to 50 cases? At present, this is an unresolved research question and a direction for future research.

References

Agresti, A. (1996). *An introduction to categorical data analysis*. New York: John Wiley & Sons.

DiBello, L. and Crone, C. (2001, July). Enhanced Score Reporting on A National Standardized Test. Paper presented at the International meeting of the Psychometric Society, Osaka, Japan.

DiBello, L. (2002, April). Skill-based scoring models for the PSAT/NMSQT. In Kristen L. Huff (Organizer), *Reporting more than scores: Skills-based scoring of a national test.* Symposium conducted at the meeting of the National Council of Measurement in Education, New Orleans, LA.

Dorans, N. J. & Holland, P. W. (1993). DIF detection and description: Mantle-Haenszel and standardization. In P. W. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 35-66). Hillsdale, NJ: Lawrence Erlbaum Associates.

Dorans, N. J., & Schmitt, A. J. (1991). *Constructed response and differential item functioning: A pragmatic approach.* (Research Rep. No. 91-47). Princeton, NJ: Educational Testing Service.

Gierl, M. J., Leighton, J. P. & Hunka, S. (2000). Exploring the logic of Tatsuoka's rule-space model for test development and analysis. *Educational Measurement: Issues and Practice, 19,* 34-44.

Holland, P. W. (1989). A note on the covariance of the Mantel-Haenszel log-odds estimator and the sample marginal rates. *Biometrics, 45,* 1009-1015. In P. W. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 35-66). Hillsdale, NJ: Lawrence Erlbaum Associates.

Keppel, G. (1991). *Design and analysis: A researcher's handbook* (3rd ed.). Upper Saddle River, NJ: Prentice Hall.

Kirk, R. E. (1995). *Experimental design: Procedures for the behavioral sciences* (3rd edition). Monteray, CA: Brooks/Cole.

Lawrence, I. (1989^{a}). *Materials for Summary of Answers Presentation on March 15, 1989.* Unpublished memorandum dated March 13, 1989.

Lawrence, I. (1989^{b}). *Proposed Graphs for the 1989 Summary of Answers Report.* Unpublished memorandum dated March 28, 1989.

Lyu, C. F., Dorans, N. J., & Ransay, J. O. (1995). *Smoothed standardization assessment of testlet level DIF on a math free-response item type.* (Research Rep. No. 95-38). Princeton, NJ: Educational Testing Service.

Mantel, N. & Haenszel, W. (1959). Statistical aspects of the analysis of data from retrospective studies of disease. *Journal of the National Cancer Institute, 22,* 719-748. In P. W. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 35-66). Hillsdale, NJ: Lawrence Erlbaum Associates.

Penfield, R. D., & Lam, T. C. M. (2000). Assessing differential item functioning in performance assessment: Review and Recommendations. *Educational Measurement: Issues and Practice, 19,* 5-15.

Ramsay, J. O. (1995). *TestGraf: A program for the graphical analysis of multiple choice test and questionnaire data.* [computer software].

Tatsuoka, K. K. (1983). Rule space: An approach for dealing with misconceptions based on item response theory. *Journal of Educational Measurement, 20,* 345-354.

Tatsuoka, K. K. (1997). Use of generalized person-fit indices for statistical pattern classification. *Journal of Applied Educational Measurement, 9,* 65-75.

Tatsuoka, K. K., Linn, R. L., Tatsuoka, M. M. & Yamamoto, K. (1988). Differential item functioning resulting from the use of different solution strategies. *Journal of Educational Measurement, 25,* 301-319.

Tatsuoka, K. K., & Tatsuoka, M. M. (1992). *A psychometrically sound cognitive diagnostic model: Effect of remediation as empirical validity.* Technical Report, RR-92-38-ONR. Princeton, NJ: Educational Testing Service.

The College Board. (1990). *PSAT/NMSQT®: Guide to the summary of answers.* New York, the College Board: Author.

Vogt, P. W. (1999). *Dictionary of statistics and methodology: A nontechnical guide for the social sciences*. Newbury Park, CA: Sage.

Zieky, M. (1993). Practical questions in the use of DIF statistics in test development. In P. W. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 35-66). Hillsdale, NJ: Lawrence Erlbaum Associates.

Authors' Note

Glenn B. Milewski, Department of Psychology, Fordham University.

Patricia A. Baron, Educational Testing Service.

We are grateful to Carolyn Crone, Neil Dorans and Charles Lewis for helpful discussions on statistical methods; Jeannine Gilsdorf, Shelby Hammond, Thanos Patelis, Ellen Sawtell and Jennifer Zanna for assistance with issues related to computing and data integrity; and, Lou DiBello, Maureen Ewing, Amy Elizabeth Schmidt, and Cathy Wendler for their helpful comments on a previous version of this paper.

Correspondence regarding this paper should be addressed to the first author and sent to Fordham University, Dealy Hall, Department of Psychology, 441 East Fordham Road, Bronx NY, 10458. E-mail correspondence can be addressed to milewski@fordham.edu

Table 1

Reference and Focal Group Descriptive Statistics

Group	Number of Examinees	Mean	SD	Percentages									
				Race-Ethnicity						Sex		Grade	
				American Indian	Asian American	African American	Hispanic	White	Other	Male	Female	10th	11th
Reference	180,643	46.6	10.9	0.01	0.06	0.14	0.10	0.68	0.03	0.45	0.55	0.42	0.58
Focal A	19,783	51.8	9.9	0.00	0.06	0.02	0.01	0.89	0.01	0.42	0.58	0.12	0.88
Focal B	24	45.3	13.0	0.00	0.28	0.06	0.11	0.44	0.11	0.50	0.50	--	--
Focal C	403	32.0	7.4	0.01	0.00	0.22	0.77	0.00	0.00	0.43	0.57	--	--
Focal D	24	32.3	5.2	0.00	0.00	0.04	0.04	0.92	0.00	0.71	0.29	--	--
Focal E	25	64.0	7.3	0.00	0.04	0.00	0.00	0.96	0.00	0.56	0.44	--	--
Focal F	369	45.5	13.7	0.01	0.10	0.22	0.01	0.63	0.03	0.56	0.44	--	--
Focal G	25	55.8	12.0	0.00	0.17	0.00	0.04	0.19	0.00	0.52	0.48	--	--
Focal H	496	54.3	14.2	0.00	0.48	0.05	0.04	0.40	0.04	0.50	0.50	--	--
Focal I	396	72.3	4.8	0.01	0.27	0.02	0.04	0.64	0.04	0.58	0.42	--	--

Note. Focal groups B - I are comprised only of 11th graders.

Table 2

DIF Statistics on Math Skills for Focal Group A (state sample)

	Differential Item Functioning (DIF) Statistics								
	Binary Skill Classifications				Continuous Posterior Probabilities				
							ANCOVA		
Math Skill Number	MH D-DIF	Cat.[a]	STD P-DIF	Cat.[b]	Standardized Mean DIF	Cat.[b]	Unstandardized b	Cat.[a]	95% Confidence Interval
1	0.24		0.012		0.00		-0.03	b	-0.05 - -0.01
2	0.51		0.014		0.00		-0.03	b	-0.04 - -0.01
3	-0.18		-0.010		0.00		0.01	b	0.00 - 0.02
4	0.33		0.014		0.00		0.00		-0.01 - 0.01
5	-0.07		-0.003		0.00		-0.01		-0.02 - 0.00
6	0.01		0.000		0.00		-0.12	b	-0.15 - -0.09
7	0.34		0.016		0.00		0.02	b	0.01 - 0.03
8	-0.03		-0.001		0.00		0.00		-0.01 - 0.00
9	0.12		0.006		0.01		0.01		0.00 - 0.02
10	0.06		0.006		0.00		0.00		-0.01 - 0.02
11	-0.04		-0.001		0.00		0.00		0.00 - 0.00
13	0.62		0.038		0.00		0.03	b	0.02 - 0.04
14	-0.07		-0.003		0.00		0.01		-0.02 - 0.00
15	0.01		0.000		0.00		0.02	b	0.01 - 0.03
16	0.02		0.001		0.00		0.01		0.00 - 0.02

Note. Cat. = category.

[a]For MH D-DIF and ANCOVA, the letters a, b, and c refer to negligible, moderate, and large DIF, respectively; blank = a.

[b]For STD P-DIF and Standardized Mean DIF, the asterisk (*) refers to a DIF value that is unusually high.

Table 3

DIF Statistics on Math Skills for Focal Group B (N = 24, low mean, high s.d.)

Differential Item Functioning (DIF) Statistics									
	Binary Skill Classifications				Continuous Posterior Probabilities				
							ANCOVA		
Math Skill Number	MH D-DIF	Cat.[a]	STD P-DIF	Cat.[b]	Standardized Mean DIF	Cat.[b]	Unstandardized b	Cat.[a]	95% Confidence Interval
1	-1.56		-0.073		0.01		0.07		-0.39 - 0.54
2	-0.33		-0.007		-0.01		-0.05		-0.33 - 0.23
3	0.95		0.063		-0.01		0.02		-0.16 - 0.20
4	2.00		0.045		-0.03		0.10		-0.20 - 0.39
5	1.16		0.034		-0.02		0.13		-0.19 - 0.44
6	-0.02		0.000		0.00		0.32		-0.50 - 1.13
7	-2.47		-0.089		0.01		-0.23		-0.50 - 0.03
8	-1.81		-0.045		-0.01		-0.01		-0.24 - 0.21
9	-1.84		-0.122	*	-0.02		-0.08		-0.37 - 0.22
10	1.28		0.092		0.03		-0.40		-0.83 - 0.02
11	-1.04		-0.016		-0.02		0.05		0.13 - 0.24
13	-0.90		-0.043		0.02		-0.22		-0.55 - 0.11
14	-2.24		-0.131	*	-0.01		-0.07		-0.32 - 0.18
15	0.45	b	0.030		-0.01		0.01		-0.29 - 0.30
16	-1.60		-0.126	*	0.00		-0.05		-0.24 - 0.15

Note. Cat. = category.

[a] For MH D-DIF and ANCOVA, the letters a, b, and c refer to negligible, moderate, and large DIF, respectively; blank = a.

[b] For STD P-DIF and Standardized Mean DIF, the asterisk (*) refers to a DIF value that is unusually high.

Table 4

DIF Statistics on Math Skills for Focal Group C (N = 403, low mean, low s.d.)

Differential Item Functioning (DIF) Statistics									
	Binary Skill Classifications				Continuous Posterior Probabilities				
							ANCOVA		
Math Skill Number	MH D-DIF	Cat.[a]	STD P-DIF	Cat.[b]	Standardized Mean DIF	Cat.[b]	Unstandardized b	Cat.[a]	95% Confidence Interval
1	-0.61		-0.012		-0.03		-0.19	b	-0.30 - -0.08
2	0.79		0.011		0.01		-0.46	b	-0.53 - -0.39
3	0.89		0.064		0.01		-0.18	b	-0.23 - -0.14
4	-0.84		-0.012		0.00		-0.39	b	-0.46 - -0.31
5	0.84		0.013		-0.01		-0.38	b	-0.46 - -0.31
6	-0.08		-0.001		-0.02		1.77	c	1.57 - 1.97
7	0.24		0.008		0.01		-0.32	b	-0.38 - -0.25
8	0.41		0.008		0.00		-0.26	b	0.32 - -0.21
9	0.15		0.004		-0.01		-0.37		-0.45 - -0.30
10	0.66		0.039		0.00		-0.04		-0.15 - 0.06
11	0.42		0.012		-0.01		0.06	b	0.01 - 0.10
13	-0.88		-0.030		-0.01		-0.49	b	-0.57 - -0.41
14	-0.23		-0.007		0.02		-0.48	b	-0.54 - -0.42
15	-1.25	b	-0.062		-0.04		-0.15	b	-0.23 - -0.08
16	0.32		0.032		0.00		-0.08	b	-0.12 - -0.03

Note. Cat. = category.

[a] For MH D-DIF and ANCOVA, the letters a, b, and c refer to negligible, moderate, and large DIF, respectively; blank = a.

[b] For STD P-DIF and Standardized Mean DIF, the asterisk (*) refers to a DIF value that is unusually high.

Table 5

DIF Statistics on Math Skills for Focal Group D (N = 24, low mean, low s.d.)

Differential Item Functioning (DIF) Statistics									
	Binary Skill Classifications				Continuous Posterior Probabilities				
							ANCOVA		
Math Skill Number	MH D-DIF[a]	Cat.[b]	STD P-DIF	Cat.[c]	Standardized Mean DIF	Cat.[c]	Unstandardized b	Cat.[b]	95% Confidence Interval
1	1.50		0.038		-0.01		-0.19		-0.66 - 0.27
2	4.17		0.059		0.02		-0.48	b	-0.76 - -0.20
3	2.07		0.158	*	0.01		-0.18		-0.36 - 0.00
4	na		-0.032		0.00		-0.29		-0.59 - 0.01
5	na		-0.022		-0.06	*	0.06		-0.26 - 0.37
6	na		-0.012		-0.01		1.67	c	0.86 - 2.48
7	1.31		0.045		0.03		-0.45	b	-0.72 - -0.19
8	na		-0.032		0.00		-0.14		-0.36 - 0.08
9	1.07		0.030		-0.01		-0.32	b	-0.62 - -0.02
10	3.32	c	0.274	*	0.08	*	-0.56	b	-0.99 - -0.14
11	2.77		0.109	*	0.01		-0.08		-0.26 - 0.11
13	0.56		0.020		0.00		-0.51	b	-0.84 - -0.17
14	-1.49		-0.035		0.00		-0.39	b	-0.65 - -0.14
15	-1.69		-0.066		-0.06	*	0.00		-0.30 - 0.30
16	0.15		0.015		0.01		-0.15		-0.35 - 0.04

Note. Cat. = category.

[a] The label 'na' pertains to circumstances when MH D-DIF could not be calculated because of a lack of variation in binary skill classification (focal group values equal to either '1' or '0').

[b] For MH D-DIF and ANCOVA, the letters a, b, and c refer to negligible, moderate, and large DIF, respectively; blank = a.

[c] For STD P-DIF and Standardized Mean DIF, the asterisk (*) refers to a DIF value that is unusually high.

Table 6

DIF Statistics on Math Skills for Focal Group E (N = 25, high mean, low s.d.)

Differential Item Functioning (DIF) Statistics									
	Binary Skill Classifications				Continuous Posterior Probabilities				
							ANCOVA		
Math Skill Number	MH D-DIF[a]	Cat.[b]	STD P-DIF	Cat.[c]	Standardized Mean DIF	Cat.[c]	Unstandardized b	Cat.[b]	95% Confidence Interval
1	0.73		0.006		0.00		0.00		-0.45 - 0.46
2	na		0.028		-0.01		-0.31		-0.59 - -0.04
3	-1.40		-0.022		0.02		-0.12		-0.29 - 0.06
4	na		0.042		0.00		-0.20		-0.50 - 0.09
5	-0.55		-0.002		0.00		-0.22		-0.53 - 0.08
6	-0.41		-0.019		0.00		2.48	c	1.68 - 3.27
7	-0.15		-0.002		0.01		-0.65	c	-0.91 - -0.39
8	na		0.037		0.02		-0.25	b	-0.46 - -0.03
9	-2.49		-0.036		-0.01		-0.61	b	-0.90 - -0.32
10	0.32		0.012		0.00		-0.33		-0.74 - 0.09
11	-0.29		-0.001		0.03		-0.01		-0.20 - 0.17
13	-1.92		-0.045		-0.05	*	-0.31		-0.64 - 0.02
14	0.57		0.004		0.04		-0.49	b	-0.73 - -0.24
15	-2.25		-0.058		-0.01		-0.17		-0.46 - 0.12
16	na		0.029		0.03		-0.14		-0.33 - 0.05

Note. Cat. = category.

[a] The label 'na' pertains to circumstances when MH D-DIF could not be calculated because of a lack of variation in binary skill classification (focal group values equal to either '1' or '0').

[b] For MH D-DIF and ANCOVA, the letters a, b, and c refer to negligible, moderate, and large DIF, respectively; blank = a.

[c] For STD P-DIF and Standardized Mean DIF, the asterisk (*) refers to a DIF value that is unusually high.

Table 7

DIF Statistics on Math Skills for Focal Group F (N = 369, low mean, high s.d.)

	Differential Item Functioning (DIF) Statistics								
	Binary Skill Classifications				Continuous Posterior Probabilities				
							ANCOVA		
Math Skill Number	MH D-DIF	Cat.[a]	STD P-DIF	Cat.[b]	Standardized Mean DIF	Cat.[b]	Unstandardized b	Cat.[a]	95% Confidence Interval
1	-1.03	b	-0.039		-0.01		-0.06		-0.18 - 0.06
2	-0.87		-0.017		-0.01		-0.10		-0.17 - 0.03
3	-0.78		-0.044		0.00		-0.11	b	-0.16 - -0.07
4	-0.43		-0.012		-0.01		-0.14	b	-0.21 - -0.06
5	-0.64		-0.016		0.00		-0.14	b	-0.22 - -0.06
6	-0.93		-0.033		-0.01		0.87	c	0.66 - 1.07
7	-1.43	b	-0.050		-0.01		-0.12	b	-0.19 - -0.05
8	0.25		0.007		0.00		-0.11	b	-0.17 - -0.05
9	-0.25		-0.009		-0.01		-0.15	b	-0.22 - -0.07
10	0.09		0.006		-0.01		-0.02		-0.13 - 0.09
11	-0.47		-0.012		-0.01		-0.01		-0.05 - 0.04
13	0.52		0.026		0.00		-0.28	b	-0.37 - -0.20
14	-0.84		-0.033		-0.01		-0.16	b	-0.22 - -0.09
15	-0.35		-0.020		0.00		-0.21	b	-0.29 - -0.14
16	0.22		0.015		0.00		-0.10	b	-0.15 - -0.05

Note. Cat. = category.

[a] For MH D-DIF and ANCOVA, the letters a, b, and c refer to negligible, moderate, and large DIF, respectively; blank = a.

[b] For STD P-DIF and Standardized Mean DIF, the asterisk (*) refers to a DIF value that is unusually high.

Table 8

DIF Statistics on Math Skills for Focal Group G (N = 25, high mean, high s.d.)

Differential Item Functioning (DIF) Statistics									
	Binary Skill Classifications				Continuous Posterior Probabilities				
							ANCOVA		
Math Skill Number	MH D-DIF	Cat.[a]	STD P-DIF	Cat.[b]	Standardized Mean DIF	Cat.[b]	Unstandardized b	Cat.[a]	95% Confidence Interval
1	0.33		0.012		0.03		0.12		-0.34 - 0.580
2	1.85		0.063		0.00		0.23		-0.05 - 0.500
3	1.60		0.065		0.00		-0.12		-0.30 - 0.050
4	-0.09		-0.003		0.01		-0.01		-0.31 - 0.280
5	-2.07		-0.054		0.03		-0.04		-0.35 - 0.260
6	-1.09		-0.043		0.00		0.93	c	0.14 - 1.730
7	-0.87		-0.034		0.00		-0.16		-0.41 - 0.100
8	3.50		0.065		0.01		-0.04		-0.26 - 0.180
9	1.24		0.045		0.02		-0.24		-0.53 - 0.050
10	-2.95	b	-0.186	*	-0.04		0.12		-0.30 - 0.540
11	3.73		0.070		-0.01		-0.10		-0.28 - 0.080
13	-1.26		-0.047		0.02		-0.21		-0.53 - 0.120
14	1.40		0.044		0.01		-0.17		-0.42 - 0.080
15	0.61		0.028		-0.02		-0.12		-0.41 - 0.170
16	-1.56		-0.055		-0.01		-0.08		-0.27 - 0.110

Note. Cat. = category.

[a] For MH D-DIF and ANCOVA, the letters a, b, and c refer to negligible, moderate, and large DIF, respectively; blank = a.

[b] For STD P-DIF and Standardized Mean DIF, the asterisk (*) refers to a DIF value that is unusually high.

Table 9

DIF Statistics on Math Skills for Focal Group H (N = 496, high mean, high s.d.)

Differential Item Functioning (DIF) Statistics									
	Binary Skill Classifications				Continuous Posterior Probabilities				
							ANCOVA		
Math Skill Number	MH D-DIF	Cat.[a]	STD P-DIF	Cat.[b]	Standardized Mean DIF	Cat.[b]	Unstandardized b	Cat.[a]	95% Confidence Interval
1	-0.14		-0.005		0.01		-0.340		-0.44 - -0.24
2	-0.16		-0.003		0.00		-0.22		-0.29 - -0.16
3	-0.44		-0.022		-0.01		-0.38		-0.42 - -0.34
4	-0.38		-0.014		-0.01		-0.12		-0.18 - -0.05
5	-0.84		-0.023		0.00		-0.23		-0.30 - -0.16
6	1.03	b	0.039		0.00		0.75		0.57 - 0.93
7	0.04		0.001		0.00		-0.58		-0.64 - -0.52
8	-1.07	b	-0.032		-0.01		-0.17		-0.22 - -0.12
9	-0.24		-0.008		0.01		-0.56		-0.62 - -0.49
10	0.31		0.018		0.01		-0.54		-0.64 - -0.45
11	-0.79		-0.018		0.00		-0.28		-0.32 - -0.24
13	-0.15		-0.007		0.00		-0.57		-0.64 - -0.49
14	-0.74		-0.030		-0.01		-0.37		-0.43 - -0.31
15	-0.33		-0.017		0.00		-0.44		-0.50 - -0.37
16	0.06		0.003		0.00		-0.43		-0.47 - -0.38

Note. Cat. = category.

[a] For MH D-DIF and ANCOVA, the letters a, b, and c refer to negligible, moderate, and large DIF, respectively; blank = a.

[b] For STD P-DIF and Standardized Mean DIF, the asterisk (*) refers to a DIF value that is unusually high.

Table 10

DIF Statistics on Math Skills for Focal Group 1 (N = 396, high mean, low s.d.)

Differential Item Functioning (DIF) Statistics									
	Binary Skill Classifications				Continuous Posterior Probabilities				
							ANCOVA		
Math Skill Number	MH D-DIF[a]	Cat.[b]	STD P-DIF	Cat.[c]	Standardized Mean DIF	Cat.[c]	Unstandardized b	Cat.[b]	95% Confidence Interval
1	na		0.001		0.00		-0.53	b	-0.65 - -0.41
2	0.00		0.000		0.00		-0.63	b	-0.70 - -0.56
3	-0.46		-0.001		0.00		-1.61	c	-1.65 - -1.56
4	0.69		0.001		0.00		-0.74	c	-0.82 - -0.67
5	na		0.000		0.00		-0.85	c	-0.93 - 0.77
6	0.29		0.003		0.00		3.59	c	3.39 - 3.79
7	na		0.000		0.00		-2.30	c	-2.37 - -2.23
8	na		0.003		0.00		-1.14	c	-1.20 - -1.09
9	na		0.001		0.00		-2.26	c	-2.34 - -2.19
10	2.83		0.005		0.00		-2.27	c	-2.38 - -2.17
11	na		0.001		0.00		-1.30	c	-1.35 - -1.26
13	1.41		0.004		0.00		-2.23	c	-2.31 - -2.15
14	na		0.003		0.01		-1.88	c	-1.94 - -1.81
15	1.44		0.002		0.00		-1.68	c	-1.75 - -1.61
16	-1.75		-0.003		0.00		-1.83	c	-1.88 - -1.78

Note. Cat. = category.

[a] The label 'na' pertains to circumstances when MH D-DIF could not be calculated because of a lack of variation in binary skill classification (focal group values equal to either '1' or '0').

[b] For MH D-DIF and ANCOVA, the letters a, b, and c refer to negligible, moderate, and large DIF, respectively; blank = a.

[c] For STD P-DIF and Standardized Mean DIF, the asterisk (*) refers to a DIF value that is unusually high.

Appendix A

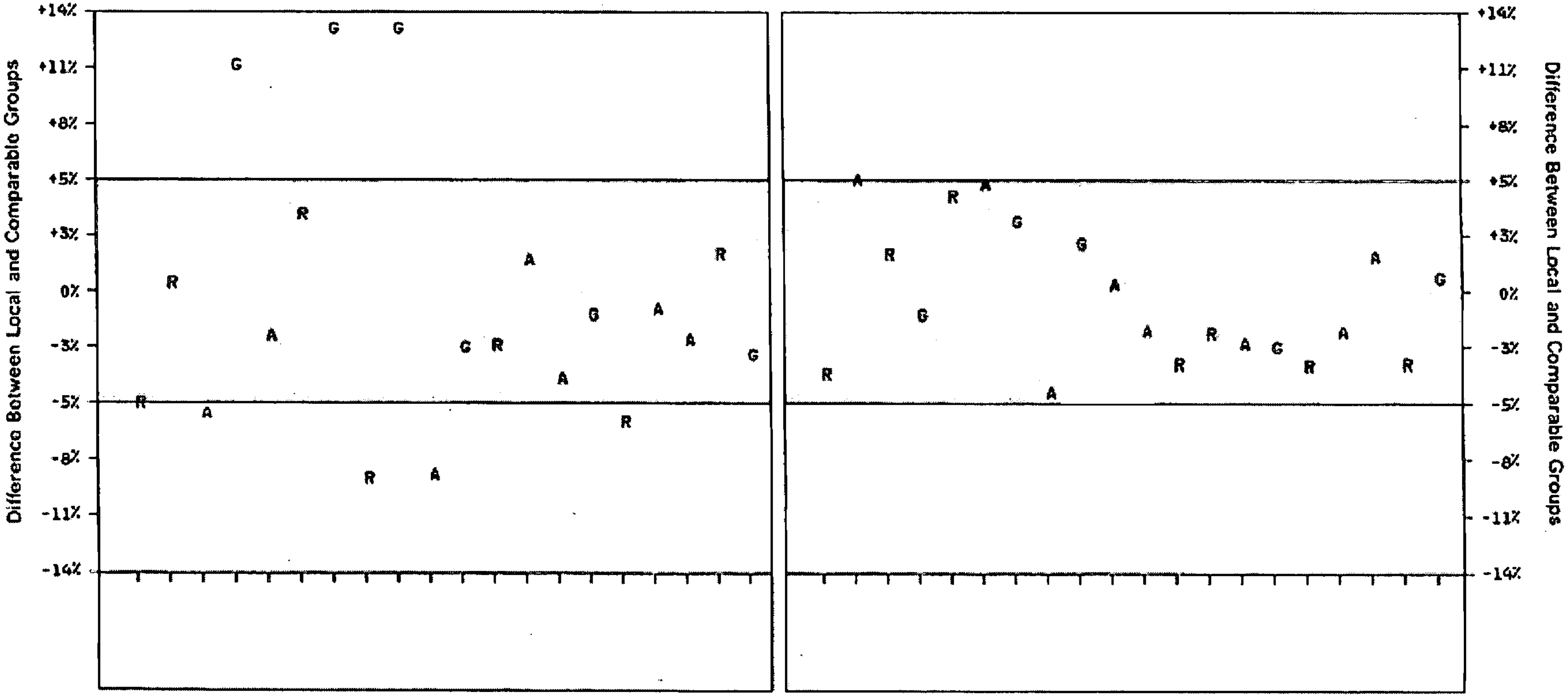
1. Using basic concepts and operation in arithmetic problem solving.
2. Understanding geometry and coordinate geometry.
3. Understanding number properties and relationships.
4. Dealing with probability, basic statistics, charts, and graphs.
5. Creating figures or algebraic equations to help solve problems.
6. Applying rules in algebra and geometry.
7. Making connections among mathematical topics.
8. Considering different cases to solve problems.
9. Organizing and managing information to help solve multistep problems.
10. Recognizing patterns and equivalent forms.
11. Using logical reasoning.
12. Searching for a solution by trying and adapting a variety of strategies.
13. Solving problems that appear unfamiliar.
14. Recognizing logical key words.
15. Using answer choices to help solve the problem.
16. Deciding when a problem doesn't provide enough information to determine a single solution.

Preliminary SAT/National Merit Scholarship Qualifying Test
SATURDAY, October 20, 2001

Code: 99999		**Grade:**	**11**
	ACORN HIGH SCHOOL	**Number of Students:**	**200**

These graphs show how your students did when matched to the comparable group.
Your student's skill mastery:

- ABOVE the band was BETTER,
- WITHIN the band was the SAME, and
- BELOW the band was WORSE.



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

ERIC®

TM034269

REPRODUCTION RELEASE

(Specific Document)

I. DOCUMENT IDENTIFICATION:

Title: EXTENDING DIF METHODS TO INFORM AGGREGATE REPORTS ON COGNITIVE SKILLS

Author(s): GLENN B. MILEWSKI & PATRICIA A. BARON

Corporate Source: NATIONAL COUNCIL OF MEASUREMENT IN EDUCATION (NCME)

Publication Date: APRIL 2002

II. REPRODUCTION RELEASE:

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

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

The sample sticker shown below will be affixed to all Level 1 documents

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL HAS BEEN GRANTED BY

Sample

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

1

Level 1

☒

Check here for Level 1 release, permitting reproduction and dissemination in microfiche or other ERIC archival media (e.g., electronic) and paper copy.

The sample sticker shown below will be affixed to all Level 2A documents

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL IN MICROFICHE, AND IN ELECTRONIC MEDIA FOR ERIC COLLECTION SUBSCRIBERS ONLY, HAS BEEN GRANTED BY

Sample

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

2A

Level 2A

□

Check here for Level 2A release, permitting reproduction and dissemination in microfiche and in electronic media for ERIC archival collection subscribers only

The sample sticker shown below will be affixed to all Level 2B documents

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL IN MICROFICHE ONLY HAS BEEN GRANTED BY

Sample

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

2B

Level 2B

□

Check here for Level 2B release, permitting reproduction and dissemination in microfiche only

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

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

Sign here,→ please

Signature: Glenn	Printed Name/Position/Title: GLENN MILEWSKI ASSISTANT RSCH SCIENTIST
Organization/Address: THE COLLEGE BOARD, 45 COLUMBUS AVE., NEW YORK, NY 10023-6992	Telephone: (212) 649-8495 FAX: (212) 649-8427
	E-Mail Address: GMILEWSKI@COLLEGEBOARD.ORG Date: 6-7-2002

(over)

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

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

Publisher/Distributor:

Address:

Price:

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

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

Name:

Address:

V. WHERE TO SEND THIS FORM:

Send this form to the following ERIC Clearinghouse:

University of Maryland
ERIC Clearinghouse on Assessment and Evaluation
1129 Shriver Laboratory
College Park, MD 20742
Attn: Acquisitions

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

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

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