

DOCUMENT RESUME

ED 335 366

TM 016 960

AUTHOR Shealy, Robin; Stout, William  
 TITLE A Procedure To Detect Test Bias Present Simultaneously in Several Items.  
 INSTITUTION Illinois Univ., Urbana. Dept. of Statistics.  
 SPONS AGENCY Office of Naval Research, Arlington, VA. Cognitive and Neural Sciences Div.  
 PUB DATE 25 Apr 91  
 NOTE 42p.  
 PUB TYPE Reports - Evaluative/Feasibility (142)

EDRS PRICE MF01/PC02 Plus Postage.  
 DESCRIPTORS \*Ability; Computer Simulation; \*Equations (Mathematics); \*Item Bias; Item Response Theory; \*Mathematical Models; \*Testing Problems  
 IDENTIFIERS Ability Estimates; Mantel Haenszel Procedure; \*Simultaneous Item Bias Procedure

ABSTRACT

A statistical procedure is presented that is designed to test for unidirectional test bias existing simultaneously in several items of an ability test, based on the assumption that test bias is incipient within the two groups' ability differences. The proposed procedure--Simultaneous Item Bias (SIB)--is based on a multidimensional item response theory (IRT) approach. SIB statistically tests for bias in one or more items at a time, and is corrected for the inflation (or deflation) of the test statistic due to target ability difference, a valid group difference that is conceptually independent of psychological test bias. The correction plays the same role as does the practice of including the single studied item in the matching criterion score in the Mantel-Haenszel (MH) procedure that is adapted for test responses. It is shown through the initial portion of an extensive simulation study in progress with 84 cases that, with the correction in place, the procedure performs as well as does the MH procedure in many cases when there is a single biased item, and it performs well in the case of multiple item test bias. Twelve tables present data from the simulation, and four graphs illustrate study findings. A 20-item list of references is included. (Author/SLD)

\*\*\*\*\*  
 \* Reproductions supplied by EDRS are the best that can be made \*  
 \* from the original document. \*  
 \*\*\*\*\*

U.S. DEPARTMENT OF EDUCATION  
Office of Educational Research and Improvement  
EDUCATIONAL RESOURCES INFORMATION  
CENTER (ERIC)

- This document has been reproduced as received from the person or organization originating it.
  - Minor changes have been made to improve reproduction quality.
- 
- Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

**BEST COPY AVAILABLE**

REPORT DOCUMENTATION PAGE				Form Approved OMB No. 0704-0188	
1a. REPORT SECURITY CLASSIFICATION <b>Unclassified</b>			1b. RESTRICTIVE MARKINGS		
2a. SECURITY CLASSIFICATION AUTHORITY		3. DISTRIBUTION / AVAILABILITY OF REPORT <b>Approved for public release; distribution unlimited</b>			
2b. DECLASSIFICATION / DOWNGRADING SCHEDULE		4. PERFORMING ORGANIZATION REPORT NUMBER(S) <b>1991 - #3</b>			
6a. NAME OF PERFORMING ORGANIZATION <b>University of Illinois Department of Statistics</b>		6b. OFFICE SYMBOL <i>(if applicable)</i>	7a. NAME OF MONITORING ORGANIZATION <b>Cognitive Science Program Office of Naval Research (Code 1142 CS)</b>		
6c. ADDRESS (City, State, and ZIP Code) <b>101 Illini Hall 725 S. Wright St. Champaign, IL 61820</b>		7b. ADDRESS (City, State, and ZIP Code) <b>800 N. Quincy Arlington, VA 22217-5000</b>			
8a. NAME OF FUNDING / SPONSORING ORGANIZATION		8b. OFFICE SYMBOL <i>(if applicable)</i>	9. PROCUREMENT INSTRUMENT IDENTIFICATION NUMBER <b>N00014-90-J-1940</b>		
8c. ADDRESS (City, State, and ZIP Code) <b>101 Illini Hall 725 S. Wright St. Champaign, IL 61820</b>		10. SOURCE OF FUNDING NUMBERS			
		PROGRAM ELEMENT NO.	PROJECT NO.	TASK NO.	WORK UNIT ACCESSION NO.
		<b>61153N</b>	<b>RR04204</b>	<b>RR04204-01</b>	<b>4421-548</b>
11. TITLE (Include Security Classification) <b>A Procedure to Detect Item Bias Present Simultaneously in Several Items</b>					
12. PERSONAL AUTHOR(S) <b>Robin Shealy and William Stout</b>					
13a. TYPE OF REPORT <b>technical</b>		13b. TIME COVERED FROM <b>1988</b> TO <b>1991</b>		14. DATE OF REPORT (Year, Month, Day) <b>April 25, 1991</b>	15. PAGE COUNT <b>35</b>
16. SUPPLEMENTARY NOTATION <b>Software to carry out the procedure is available from the authors</b>					
17. COSATI CODES			18. SUBJECT TERMS (Continue on reverse if necessary and identify by block number)		
FIELD	GROUP	SUB-GROUP	<b>See reverse</b>		
19. ABSTRACT (Continue on reverse if necessary and identify by block number)  <b>See reverse</b>					
20. DISTRIBUTION / AVAILABILITY OF ABSTRACT <input checked="" type="checkbox"/> UNCLASSIFIED/UNLIMITED <input type="checkbox"/> SAME AS RPT. <input type="checkbox"/> DTIC USERS			21. ABSTRACT SECURITY CLASSIFICATION		
22a. NAME OF RESPONSIBLE INDIVIDUAL <b>Dr. Charles E. Davis</b>			22b. TELEPHONE (Include Area Code) <b>(703) 696-4046</b>	22c. OFFICE SYMBOL <b>ONR-1142-CS</b>	

## **ABSTRACT**

This paper presents a statistical procedure (denoted by SIB) designed to test for unidirectional test bias existing simultaneously in several items of an ability test. It was argued in Shealy and Stout (1991) that in order to model such bias with an IRT model, a multidimensional model is necessary. The proposed procedure, based on this multidimensional IRT modeling approach, statistically tests for bias in one or more items at a time and is corrected for the inflation (or deflation) of the test statistic due to target ability difference, a valid group difference that is conceptually independent of psychological test bias. The correction plays the same role as the practice of including the single studied item in the "matching criterion" score in the Mantel-Haenszel (MH) procedure adapted for test responses by Holland and Thayer (1988). It is shown through the initial portion of an extensive simulation study underway (Shealy (1991)) that, with the correction in place, the procedure performs as well as the MH procedure in many cases when there is a single biased item, and performs well in the case of multiple item test bias.

**Key Words:** item bias, test bias, DIF, latent trait theory, item response theory, target ability, valid subtest, nuisance determinants, potential for bias, expressed bias, unidirectional test bias, bidirectional test bias, SIB, Mantel-Haenszel.

## INTRODUCTION

The purpose of this paper is to present a statistical procedure (denoted by SIB for simultaneous item bias) for detecting bias present in one or more test items of a standardized ability test. The procedure is based on the multidimensional item response theory (IRT) model of test bias presented in Shealy and Stout (1991). By "test bias" we mean a formalization of the intuitive idea that a test is less valid for one group of examinees than for another group in its attempt to assess examinee differences in a prescribed latent trait, such as mathematics ability. Test bias is conceptualized herein as the result of individually-biased items acting in concert through a test scoring method, such as number correct, to produce a biased test.

Two distinct features of this conceptualization of bias are as follows. First, it provides a mechanism for explaining how several individually-biased items can combine through a test score to exhibit a coherent and major biasing influence at the test level. In particular, this can be true even if each individual item displays only a minor amount of item bias. For example, word problems on a mathematics test that are too dependent on sophisticated written English comprehension could *combine* to produce pervasive test bias against English-as-a-second-language examinees. A second feature, possible because of our multidimensional modeling approach, is that the underlying psychological mechanism that produces bias is addressed. This mechanism lies in the distinction made between the ability the test is intended to measure, called the *target ability*, and other abilities influencing test performance that the test does not intend to measure, called *nuisance determinants*. Test bias will be seen to occur because of the presence of nuisance determinants possessed in differing amounts by different examinee groups. Through the presence of these nuisance determinants, bias then is expressed in one or more items.

The test bias detection procedure can simultaneously assess bias in several items, thus addressing the above two features. In contrast, most item bias procedures detailed in the literature perform tests on a single item at a time: The pseudo IRT procedure of Linn and Harnish (1981) estimates possibly group-dependent item response functions (IRFs) without the use of item parameter estimation algorithms when the sample size is too small for their use. Thissen, Steinberg, and Wainer (1988) employ marginal maximum likelihood estimation to obtain group-dependent item parameters in a 3-parameter logistic framework and use the likelihood ratio test to test the equality of the parameters across group. The Mantel-Haenszel procedure, adapted for test response data by Holland and Thayer (1988), and which is in wide use, employs the practice of using the score of the entire test instead of the score of the non-studied items as the "matching criterion" to test for item bias. Etc. Conceivably these procedures could be used once for each item in a set of items being tested for bias, and multiple comparison procedures could be employed to assess the hypothesis of the entire set being biased. However, if the amount of bias is small

in each item, a multiple comparison procedure may not pick up bias in the set of items at all. Moreover this approach cannot address underlying causal mechanisms of bias.

The novelty of our approach to detecting test bias lies not so much with its recognition of the role of nuisance determinants in the expression of test bias, but rather in its explicit use of a multidimensional model to motivate the procedure to detect it. The presence of multidimensionality of test item responses where bias is present has long been recognized in test and item bias studies: Lord (1980) states "if many of the items [in a test] are found to be seriously biased, it appears that the items are not strictly unidimensional" (p. 220). Recently, Lautenschlager and Park (1988) employed a technique of generating simulated biased item responses using a method of Ansley and Forsyth (1985), which involves using multidimensional item response functions (IRFs) and latent ability distributions to determine conditional probabilities of correct response. Kok (1988), taking a multidimensional viewpoint similar to Shealy and Stout (1991), presents a specific multidimensional IRT model for bias where the nuisance determinants are compensating abilities, contextual abilities such as language, and testwiseness.

An important issue addressed by our procedure is that a careful distinction is made between genuine test bias, often operationally embodied as DIF (Holland and Thayer (1988)) by practitioners, and non-bias differences in examinee group performance, sometimes called *impact* (see, for example, Ackerman (1991) for a careful discussion of impact as distinct from bias), that are caused by examinee group differences in target ability distributions. It is important that the latter not be mistakenly labeled as test bias. The procedure developed herein makes this distinction in its application.

## FORMULATION OF TEST BIAS

Test bias in this paper is modeled using a multidimensional item response theory (IRT) model, which is assumed to be the model behind the observed test responses. For purposes of exposition, we restrict ourselves to the case where there is a single nuisance determinant; this two-dimensional modeling approach is often realistic in practice. Extensions to multiple nuisance determinants are straightforward. For a fuller treatment of the conception of test bias, including the case of multiple nuisance determinants and item bias cancellation, in a more general framework, see Shealy and Stout (1991) and Shealy (1989).

We consider two biologically- or sociologically-defined groups, named "reference" and "focal" groups (after Holland and Thayer's (1988) naming convention). A random sample of examinees is drawn from each group, and a test of  $N$  items is administered to them. Typically it is suspected that a part of the test is biased against the focal group; this group is usually the object of the bias study. The responses to the test items from a randomly-chosen examinee are denoted  $\underline{U} = (U_1, \dots, U_N)$ , where each  $U_i$  can take on 0 or 1, according as the response to item  $i$  is incorrect or correct, respectively.

The IRT model in general is composed of two components that generate  $\underline{U}$ : (1) a  $d$ -dimensional examinee ability parameter and (2) a set of item response functions (IRFs), one for each item, which determine the probability of correct response for the items. Here we restrict the model to have  $d = 1$  or  $2$ , because we are considering a single nuisance determinant in addition to the target ability. The ability vector is  $(\theta, \eta)$  for an arbitrary examinee from either group, where  $\theta$  denotes target ability and  $\eta$  denotes the nuisance determinant. A distribution of  $(\theta, \eta)$  over the combined group of examinees is induced by choosing examinees at random; the variable for a randomly chosen examinee is denoted  $(\Theta, \eta)$ . The IRF for item  $i$  is denoted  $P_i(\theta, \eta)$ , and it is assumed that all items depend on  $\theta$ , and one or more may depend on  $\eta$ ; for those dependent only on  $\theta$ , the IRF is  $P_i(\theta)$ . It is implicitly assumed that an IRT representation for  $\underline{U}$  in terms of  $(\Theta, \eta)$  and  $\{P_i(\theta, \eta) : i = 1, \dots, N\}$  is possible; for a fuller treatment of this assumption, see Shealy (1989). In addition, it is assumed that each  $P_i(\theta, \eta)$  is increasing in  $(\theta, \eta)$  when item  $i$  is dependent on both abilities and increasing in  $\theta$  when it is dependent on  $\theta$  alone; and that each  $P_i(\theta)$  is differentiable. Finally, local independence of  $\underline{U}$  given  $(\theta, \eta)$  is assumed.

Test bias in the above-mentioned model is formulated through three components:

- (a) The *potential for bias*, if it exists, resides within the target ability/nuisance determinant distributions of the two groups being studied;
- (b) potential for bias is *expressed* in items whose responses depend on the nuisance determinant;<sup>1</sup> and

---

<sup>1</sup> We remark that Kok's (1988) formulation is also based upon (a) and (b); Kok's and our formulation were developed independently of one another.

(c) the scoring method of the test, to be viewed as an estimate of target ability, transmits expressed item biases into test bias.

Potential for test bias is explained prosaically in the following manner. After conditioning on a particular  $\theta$ , suppose that the reference group has a higher level of nuisance ability on average than the focal group. Then those reference group examinees with ability  $\theta$  would have an overall advantage over the corresponding focal group examinees when responding to items at least partially dependent on the nuisance determinants  $\eta$  (formally, because of the monotonicity of the items IRFs  $P_i(\theta, \eta)$ ). Formally, we define the potential for test bias at  $\theta$ :

**Definition 1.** *Potential for test bias* exists against the focal group at target ability level  $\theta$  with respect to  $\eta$  if  $\eta | \Theta = \theta, G = F$  is stochastically less than  $\eta | \Theta = \theta, G = R$ , where “ $G = F$ ” denotes sampling from the focal group and “ $G = R$ ” sampling from reference group. Potential for bias exists against the reference group if the converse holds.

Note that we are restricting consideration to conditional nuisance distributions  $\eta | \Theta = \theta, G = R$  and  $\eta | \Theta = \theta, G = F$  that are stochastically ordered; that is, where the two distribution functions do not intersect. Figure 1 displays two distributions that are stochastically ordered and also two distributions that are not.

place Figure 1 about here

In order for test bias to occur, it must be *expressed* in one or more items. Our definition of expressed bias for an item, when specialized to Kok’s model, is really the same as that

of Kok (1988, p. 269). It is defined in terms of a marginalization of the multidimensional IRF  $P_i(\theta, \eta)$ .

**Definition 2.** Let  $P_i(\theta, \eta)$  be the IRF for item  $i$ . The *marginal IRF* for group  $g$  ( $g = R$  or  $F$ ) with respect to target ability  $\theta$  is defined as

$$T_{i_g}(\theta) = E[P_i(\Theta, \eta) \mid \Theta = \theta, G = g]. \quad (1)$$

When  $\eta \mid \theta$  has a conditional density,  $f(\eta \mid \theta)$  say, Definition 2 translates into

$$T_{i_g}(\theta) = \int_{-\infty}^{\infty} P_i(\theta, \eta) f(\eta \mid \theta) d\eta.$$

**Definition 3.** *Expressed bias* for item  $i$  against the focal group occurs at target ability  $\theta$  if  $T_{i_F}(\theta) < T_{i_R}(\theta)$ ; it occurs against the reference group if the converse holds.

A test can consist of many items simultaneously biased by the same nuisance determinant. In this case, items can cohere and act through the prescribed test score to produce substantial bias against a particular group even if individual items display undetectably small amounts of item bias. This is the final (and novel) component of our formulation of test bias mentioned above. We consider the large class of test scores of the form

$$h(\underline{U}) \quad (2)$$

where  $h(\underline{u})$  is real valued with domain  $\underline{u} \equiv (u_1, \dots, u_N)$  such that  $u_i = 0$  or  $1$  for  $i = 1, \dots, N$  and  $h(\underline{u})$  is coordinate wise non-decreasing in  $\underline{u}$ . This class contains many of the standard scoring procedures for many standard models; for example, number correct, linear formula scoring of the form  $\sum_{i=1}^N a_i U_i$ , with  $a_i \geq 0$ , maximum likelihood estimation of ability for certain logistic models with item parameters assumed known, etc. In this paper we restrict attention to number correct as the test score; the results presented herein are easily extendable to other forms of  $h(\underline{u})$ . The key point about number correct scoring is that each item is weighted equally. Thus, if a subset of the items is suspected of bias, we should give equal weight to the items in this "studied" subtest in our attempt to quantitatively assess the amount of test bias resulting from the simultaneous influence of these items. We thus define test bias for a specified studied subtest of items as follows:

**Definition 4.** Let  $\{U_{i_1}, U_{i_2}, \dots, U_{i_b}\}$  be any subtest of items to be studied for bias from the test of concern and define

$$h(\underline{U}) = \sum_{j=1}^b U_{i_j}. \quad (3)$$

Then this studied subtest of items displays test bias against the focal group at  $\theta$  if

$$E[h(\underline{U}) \mid \Theta = \theta, G = F] < E[h(\underline{U}) \mid \Theta = \theta, G = R].$$

The subtest is biased against the reference group if the converse holds.

Finally, the components of the bias formulation can be integrated using the following theorem, adapted from Theorem 4.2 in Shealy and Stout (1991):

**Theorem 1.** Fix a target ability  $\theta$  and choose the subtest scoring method  $h(\underline{u})$  of the form (3). Assume potential for bias against the focal group at  $\theta$  holds (Definition 1). Then test bias exists against the focal group; i.e.,

$$\sum_{j=1}^b E[U_{i_j} | \Theta = \theta, G = F] < \sum_{j=1}^b E[U_{i_j} | \Theta = \theta, G = R]. \quad (4)$$

In order to test for bias of the above form, there must be an implicit assumption that a portion of the test measures only the target ability; otherwise, a conditional-on-observed score procedure to detect bias is not possible. This set of items will be denoted the *valid subtest*. The issue of the existence and identification of a valid subtest is extremely difficult to frame philosophically (it is really an issue of construct validity) and must primarily be an empirical decision based on expert opinion or data at least in part external to the test being studied; it is not dealt with here. For a fuller discussion, see Shealy and Stout (1991). For notational simplicity we denote the valid subtest to consist of first  $n < N$  items of the test, and we call the remainder of the  $N - n$  items the *studied subtest*. We note that use of a valid subtest is operationally equivalent to making use of a subset of items whose purpose is to partition examinees into “comparable” sets as is done in the MH procedure described below and other DIF procedures. Hence, the proposed use of a valid subtest in the SIB procedure can be interpreted either in the strong sense of our test bias paradigm or in the weak sense of the DIF paradigm (of matching of “comparable” examinees). Thus use of our statistical procedure for assessing bias in no way requires acceptance of our bias framework as opposed to a “comparability” framework, where no claims about “bias” are made.

Using the above conventions, the specification of test bias against the focal group at  $\theta$  becomes

$$T_F(\theta) \equiv \sum_{i=n+1}^N T_{iF}(\theta) < \sum_{i=n+1}^N T_{iR}(\theta) \equiv T_R(\theta) \quad (5)$$

because  $T_{i_g}(\theta) = E[U_i | \Theta = \theta, G = g]$  by a simple application of a standard conditioning formula to Definition 2.  $T_g(\theta)$  is called the *studied subtest response function* for group  $g$ .

### Unidirectional test bias

Test bias heretofore has been considered conditional on a single target ability; we now turn to a global perspective. If there is test bias against the same group for all  $\theta$ , then there is unidirectional bias against this group. Specifically, if

$$B(\theta) = T_R(\theta) - T_F(\theta)$$

is the level of bias against Group  $F$  at  $\theta$ , then unidirectional bias holds if either  $B(\theta) > 0$  for all  $\theta$  or  $B(\theta) < 0$  for all  $\theta$ . A strong form of unidirectional bias, termed uniform bias by Mellenbergh (1982), is the type of bias that the modified Mantel-Haenszel test statistic devised by Holland and Thayer (1988) is designed to detect. Although the Mantel-Haenszel approach is not dependent on an IRT framework, it can be put in a Rasch model IRT framework, with the single biased item having group-dependent item difficulties. Here, the bias is "uniform" in the sense that  $T_F(\theta)$  is merely  $T_R(\theta)$  shifted horizontally. Unidirectional bias is less restrictive in that  $T_g(\theta)$  does not have to be a logistic IRF, and more importantly,  $T_R(\theta)$  does not have to be  $T_F(\theta)$  shifted.

Since we are concerned with bias against the focal group, it is intuitive that a suitable theoretical unidirectional bias index is

$$\beta_U = \int_{\theta} B(\theta) f_F(\theta) d\theta \quad (6)$$

where  $f_F(\theta)$  is the probability density function of  $\Theta$  for the focal group. Equivalent indices weighted by the reference target ability distribution and the combined-group target distribution are easily conceptualized.

### THE BASIC PROCEDURE

The statistical procedure to be presented is based on (6); the hypothesis is

$$H : \beta_U = 0 \quad \text{vs.} \quad \beta_U > 0,$$

the alternative being one-sided to specifically test for bias against the focal group. The test statistic to be constructed is essentially an estimate of  $\beta_U$  normalized to have unit variance. The estimate of  $\beta_U$  is derived first.

Since test bias is analyzed using number correct on the studied subtest, set

$$Y = \sum_{i=n+1}^N U_i \quad (7)$$

to be the studied subtest score; also set  $X = \sum_{i=1}^n U_i$  to be the valid subtest score. In selecting the valid subtest score to be number correct, we follow the convention set out in Holland and Thayer (1988), among many others. Other choices would of course be possible and could improve the performance of the procedure.

The naive intuition is that examinees with the same valid subtest score are examinees of approximately equal target ability and thus such examinees are directly comparable in the assessment of bias. Thus the difference

$$\bar{Y}_{Rk} - \bar{Y}_{Fk}, \quad k = 0, \dots, n, \quad (8)$$

where  $\bar{Y}_{gk}$  is the average  $Y$  for all examinees in group  $g$  attaining valid subtest score  $X = k$ , should provide a measure of the bias against the focal group (resulting from the reference group having superior nuisance ability  $\eta$  on average). In particular, if there is no bias ( $H$  holds), then  $\bar{Y}_{Rk} - \bar{Y}_{Fk} \doteq 0$  for all  $k$  should be observed, and if there is unidirectional bias against the focal group ( $B(\theta) > 0$  for all  $\theta$ ) then  $\bar{Y}_{Rk} - \bar{Y}_{Fk} > 0$  for all  $k$ , except for statistical error, should be observed.

The above assertion needs support; it will suffice to argue that

$$\begin{aligned} E[\bar{Y}_{Rk} - \bar{Y}_{Fk}] &\doteq 0 \quad \text{for all } k \text{ if } B(\theta) = 0 \text{ for all } \theta, \text{ and} \\ E[\bar{Y}_{Rk} - \bar{Y}_{Fk}] &> 0 \quad \text{for all } k \text{ if } B(\theta) > 0 \text{ for all } \theta. \end{aligned} \quad (9)$$

For now we restrict the target ability distributions to be equal for the two groups; i.e.,  $\Theta | G = R$  and  $\Theta | G = F$  have the same distribution. It is easy to prove (following (5)) under the model presented herein that

$$E[\bar{Y}_{gk}] = E[Y | X = k, G = g] = E[T_g(\Theta) | X = k, G = g]. \quad (10)$$

Now assume that the valid subtest is long enough so that the distribution of  $\Theta | X = k, G = g$  is tightly concentrated about its mean, and hence that  $T_g(\theta)$  is locally flat within the range of  $\theta$  where the distribution of  $\Theta | X = k, G = g$  mostly resides. Then

$$\begin{aligned} E[T_g(\Theta) | X = k, G = g] &\doteq T_g(E[\Theta | X = k, G = g]) \\ &= T_g(E[\Theta | X = k]), \end{aligned} \quad (11)$$

because the two target ability distributions are equal and expectation is a linear operator. Thus, denoting  $\theta_k = E[\Theta | X = k]$ ,

$$E[\bar{Y}_{Rk} - \bar{Y}_{Fk}] \doteq B(\theta_k). \quad (12)$$

Thus (9) follows easily; the  $n + 1$  differences in (8) provide an estimate of  $B(\theta)$  at  $n + 1$  points in the  $\theta$ -domain. It is intuitive that an estimate of  $\beta_U$  is

$$\hat{\beta}_U = \sum_{k=0}^n \hat{p}_k (\bar{Y}_{Rk} - \bar{Y}_{Fk}) \quad (13)$$

where  $\hat{p}_k$  is the proportion (among focal group examinees) attaining  $X = k$ . Specifically, if  $J_{gk}$  is the number of examinees in group  $g$  attaining  $X = k$ , then  $\hat{p}_k = J_{Fk} / \sum_{k=0}^n J_{Fk}$ .

In the case where the target ability distributions are the same, then, it is straightforward that

$$E[\hat{\beta}_U] \doteq \sum_{k=0}^n p_k B(\theta_k) \doteq \beta_U \quad (14)$$

where  $p_k = P[X = k | G = F]$ . Thus the expected value of  $\hat{\beta}_U$  is a weighted difference of marginal IRFs, this weighted difference approximating  $\beta_U$ , which is a continuously weighted difference of marginal IRFs. From (14), it follows that  $E\hat{\beta}_U = 0$  if  $\beta_U = 0$ , and  $E\hat{\beta}_U > 0$  if  $\beta_U > 0$ . This suggests the standardized test statistic

$$B = \frac{\hat{\beta}_U}{\hat{\sigma}(\hat{\beta}_U)} \quad (15)$$

for testing  $H$ , where the denominator is defined as

$$\hat{\sigma}(\hat{\beta}_U) = \left( \sum_{k=0}^n \hat{p}_k^2 \left( \frac{1}{J_{Rk}} \hat{\sigma}^2(Y | k, R) + \frac{1}{J_{Fk}} \hat{\sigma}^2(Y | k, F) \right) \right)^{1/2}, \quad (16)$$

where  $\hat{\sigma}^2(Y | k, g)$  is the sample variance of the studied subtest scores of those group  $g$  examinees with valid subtest score  $k$ . A full description of the computation of the test statistic, with contingencies for exclusion of certain valid subtest scores based on inadequate examinee counts, is presented in the Appendix.  $B$  is approximately standard normal when  $\beta_U = 0$  and the target ability distributions are the same, because  $\hat{\beta}_U$  is the weighted sum of approximately normal random variables  $\bar{Y}_{Rk} - \bar{Y}_{Fk}$ ; these are approximately normal (for suitable sample sizes) by the central limit theorem (proof of asymptotic normality of  $B$  omitted).

### The regression correction for target ability difference

The presence of a difference in target ability distributions in test bias studies has been treated in various contexts in the literature. The issue of the linking of metrics across group in the estimation of IRT item parameters is one such context (see Linn, et al (1981) for an IRT item bias approach where linking of metrics is crucial). Holland and Thayer (1988) also deal with this problem by including the single studied item in the matching criterion score of the Mantel-Haenszel test; they prove that this method completely compensates for target ability difference (in their context, the distributional difference in the postulated unidimensional latent trait) when the underlying IRT model is a Rasch model. Millsap and Meredith (1989) elegantly formulate the problem in terms of a divergence of two hypotheses (a "conditional on observed score" hypothesis and a "latent trait" hypothesis), which would occur if target ability difference is present. A "conditional on observed score" procedure such as (15) in its present form is not adequate to address the separation of target ability difference from test bias; the presence of target ability difference when in fact there is no test bias present can statistically inflate  $B$ , thereby suggesting test bias actually is present. It is therefore necessary to formulate a correction for target ability difference.

To motivate the proposed correction it is necessary to show that a decomposition of the differences  $\bar{Y}_{Rk} - \bar{Y}_{Fk}$  into “test bias only” and “target ability difference only” components is possible. First we note that by similar arguments to those used in deriving (10) and (11),

$$E[\bar{Y}_{gk}] \doteq T_g(\theta_{gk}), \quad (17)$$

where  $\theta_{gk} = E[\Theta | k, g]$ . The condition  $E[\bar{Y}_{Rk} - \bar{Y}_{Fk}] \doteq 0$  requires  $\theta_{Rk} \doteq \theta_{Fk}$ , as in (11) where  $g$  was removed from the conditioning; but this may not happen if the target ability distributions are not the same, as Figure 2 suggests. Figure 2, which displays densities for four distributions, assumes that the distribution of  $\Theta | F$  is stochastically smaller than that of  $\Theta | R$ .

place figure 2 about here

Note that the (conditional) distribution of  $\Theta | k, F$  is stochastically smaller than that of  $\Theta | k, R$  for all  $k$ . The standard Bayesian calculation makes this insight rigorous. Thus,  $\theta_{Fk} < \theta_{Rk}$  for all  $k$ , and, in the absence of bias, where  $T_R(\theta) = T_F(\theta) \equiv T(\theta)$  for all  $\theta$ ,

$$E\bar{Y}_{Fk} \doteq T(\theta_{Fk}) < T(\theta_{Rk}) \doteq E\bar{Y}_{Rk}$$

( $T(\theta)$  is assumed monotone; for mild conditions giving such monotonicity, see Shealy and Stout (1991)). Thus

$$E\hat{\beta}_U \doteq \sum_{k=0}^n p_k (T(\theta_{Rk}) - T(\theta_{Fk})) > 0.$$

In the case where bias is present, we can thus decompose  $E[\hat{\beta}_U]$ :

$$\begin{aligned} E[\hat{\beta}_U] &\doteq \sum_{k=0}^n p_k (T_R(\theta_{Rk}) - T_F(\theta_{Rk})) + \sum_{k=0}^n p_k (T_F(\theta_{Rk}) - T_F(\theta_{Fk})) \\ &\doteq \sum_{k=0}^n p_k B(\theta_{Rk}) + \sum_{k=0}^n p_k T'_F(\theta_k^*)(\theta_{Rk} - \theta_{Fk}), \end{aligned} \quad (18)$$

where  $\theta_k^*$  is between  $\theta_{Rk}$  and  $\theta_{Fk}$ . ( $T_F(\theta)$  is assumed differentiable here and the mean value theorem has been applied.) The first term is due only to test bias; the second is due only to target ability difference.

This approximate decomposition argument is the motivation behind the proposed correction. Our strategy is to adjust  $\bar{Y}_{Rk}, \bar{Y}_{Fk}$  to  $\bar{Y}_{Rk}^*, \bar{Y}_{Fk}^*$  such that the inflating effect of the group differences in target ability is eliminated. The manner this is accomplished is to construct  $\bar{Y}_{Rk}^*$  and  $\bar{Y}_{Fk}^*$  so that they are estimating the studied subtest response functions  $T_R(\theta)$  and  $T_F(\theta)$  at approximately the same target ability  $\theta_k$  defined below (as opposed to two different ones, as is evident from (17)).

A natural attempt to make adjustments to  $\bar{Y}_{Rk}$  and  $\bar{Y}_{Fk}$  is to approximate  $T_R(\theta)$  and  $T_F(\theta)$  in the neighborhood of  $\theta_{Rk}$  and  $\theta_{Fk}$  by linear functions. If we assume that  $\theta_{Rk}$  and  $\theta_{Fk}$  are sufficiently close together to do this,  $T_R(\theta)$  and  $T_F(\theta)$  can be linearly interpolated at  $\theta_k = \frac{1}{2}(\theta_{Rk} + \theta_{Fk})$ :

$$T_g(\theta_k) = T_g(\theta_{gk}) + m_{gk}(\theta_k - \theta_{gk}) \quad (19)$$

where

$$m_{gk} = \frac{T_g(\theta_{g,k+1}) - T_g(\theta_{g,k-1})}{\theta_{g,k+1} - \theta_{g,k-1}},$$

however, though estimates of  $T_g(\theta_{gk})$  (namely,  $\bar{Y}_{gk}$ ) are available for all  $k$ , estimates for  $\{\theta_{gk} : k = 0, \dots, n\}$  are not. Abilities on the  $\theta$ -scale are not observable; however, *one can estimate abilities on the scale defined by the valid subtest, namely*

$$v = \bar{P}(\theta)$$

where  $\bar{P}(\theta)$  is the average of the valid subtest IRFs  $\frac{1}{n} \sum_{i=1}^n P_i(\theta)$ .  $\bar{P}(\Theta) | G = g$  is the true score for a randomly chosen group  $g$  examinee, i.e., the valid subtest true score  $\bar{P}(\Theta)$  for group  $g$ . Let

$$V_g(x) = E[\bar{P}(\Theta) | X = x, G = g], \quad (20)$$

the (theoretical) regression of true on observed (here, valid) score.  $V_g(x)$  can be easily estimated using classical true score theory, assuming that the above regression is linear or nearly so. The estimation of  $V_g(x)$  is deferred to the appendix. Denote this estimator by  $\hat{V}_g(x)$ .

At this point it is expedient to describe *three* latent scales, which must be simultaneously considered in order to understand the correction. Figure 3 delineates the three scales and should be referred to frequently.

place figure 3 about here

So, the interpolation of (19) must be transformed so as to use the easily estimable  $V_g(k)$  instead of  $\theta_{gk}$ . Through a monotonic transformation  $\bar{P}(\theta)$ ,  $V_g(k)$  and  $\theta_{gk}$  represent approximately ("approximately" because  $\bar{P}(\theta_{gk}) \doteq V_g(k)$  will be demonstrated below) the *same* ability on two different latent scales and thus for our purposes interchangeable. Note that  $s = T_g(\theta)$  defines a monotonic transformation from the fundamental latent scale to the studied subtest scale, and  $v = \bar{P}(\theta)$  defines one from the fundamental scale to the valid subtest scale.  $T_g(\theta)$  must be transformed so we can use the valid subtest scale as domain, because abilities on this scale can be estimated. Figure 4 illustrates the appropriate correspondence,

place figure 4 about here

thus defining a new transformation  $S_g(v) = T_g(\bar{P}^{-1}(v))$  from valid subtest scale to studied subtest scale, with domain  $(c, 1)$  and range  $(c, 1)$  ( $c \geq 0$  is the guessing parameter, assumed common for all items in the test).

With this transformation in hand, the correction can be performed in the following manner. First, by the same arguments as used in (10) and (11), using  $\bar{P}(\theta)$  in place of  $T_g(\theta)$  in the arguments,

$$V_g(k) \doteq \bar{P}(E[\Theta | k, g]) \equiv \bar{P}(\theta_{gk}). \quad (21)$$

So  $\bar{P}^{-1}(V_g(k)) \doteq \theta_{gk}$  by continuity; and

$$T_g(\bar{P}^{-1}(V_g(k))) \doteq T_g(\theta_{gk}),$$

also by continuity. By definition of  $S_g(v)$ , this becomes  $S_g(V_g(k)) \doteq T_g(\theta_{gk})$ , and thus by (17),

$$E\bar{Y}_{gk} \doteq S_g(V_g(k)). \quad (22)$$

Thus  $\bar{Y}_{gk}$  is a reasonable estimation of  $S_g(V_g(k))$  for each  $k$ . To transform (19) into an interpolation involving  $S_g(\cdot)$ , we assume that  $S_g(v)$  can be approximated by a linear function in a small region about  $V_g(k)$ , and that  $V_R(k)$  and  $V_F(k)$  are close enough to allow the approximation to be effective. Then, we interpolate  $S_R(V_R(k))$  and  $S_F(V_F(k))$  to their respective values at  $V_k = \frac{1}{2}(V_R(k) + V_F(k))$ :

$$S_g(V_k) \doteq S_g(V_g(k)) + m_{gk}^*(V_k - V_g(k)), \quad (23)$$

where

$$m_{gk}^* = \frac{S(V_g(k+1)) - S_g(V_g(k-1))}{V_g(k+1) - V_g(k-1)}$$

is the approximate slope of  $S_g(v)$  in the region of  $V_g(k)$  and  $V_k$ . All of the above terms on the right hand side of (23) are estimable; using  $\bar{Y}_{gk}$  to estimate  $S_g(V_g(k))$ , we define the adjusted  $\bar{Y}_{gk}^*$ :

$$\bar{Y}_{gk}^* = \bar{Y}_{gk} + \hat{M}_{gk}(\hat{V}_k - \hat{V}_g(k)) \quad (24)$$

where, recalling that the estimator  $\hat{V}_g(x)$  is given in the Appendix,

$$\hat{M}_{gk} = \frac{\bar{Y}_{g,k+1} - \bar{Y}_{g,k-1}}{\hat{V}_g(k+1) - \hat{V}_g(k-1)}$$

and define  $\hat{V}_k = \frac{1}{2}(\hat{V}_R(k) + \hat{V}_F(k))$ . Because the right hand side of equation (24) is a good estimator of the right hand side of (23),  $\bar{Y}_{gk}^*$  is thus a good estimator of  $S_g(V_k)$ . Finally,  $\bar{Y}_{gk}^*$  must be shown to be a good estimator of  $T_g(\theta)$  at the *same*  $\theta$  for both groups. By definition of  $S_g(v)$ ,  $S_g(V_k) = T_g(P^{-1}(V_k))$ . If  $\theta_{Rk}$  and  $\theta_{Fk}$  are sufficiently close together then  $\bar{P}(\theta)$  may be taken to be approximately linear in the neighborhood of  $\theta_k = (\theta_{Rk} + \theta_{Fk})/2$ . Thus, using (21) and assuming approximate linearity of  $\bar{P}$  in the neighborhood of  $\theta_k$ ,

$$\begin{aligned} V_k &= \frac{1}{2}(V_R(k) + V_F(k)) \\ &\doteq \frac{1}{2}(\bar{P}(\theta_{Rk}) + \bar{P}(\theta_{Fk})) \\ &\doteq \bar{P}(\theta_k). \end{aligned}$$

Thus, by the continuity of  $\bar{P}(\theta)$ ,

$$\theta_k \doteq \bar{P}^{-1}(V_k).$$

Hence, by the definition of  $S_g(v)$

$$S_g(V_k) = T_g(\bar{P}^{-1}(V_k)) \doteq T_g(\theta_k).$$

Thus, because  $\bar{Y}_{gk}^*$  has been shown to be a good estimator of  $S_g(V_k)$ , it is shown that  $\bar{Y}_{gk}^*$  is a good estimator of  $T_g(\theta_k)$ . Thus,  $\bar{Y}_{Rk}^* - \bar{Y}_{Fk}^*$ , as desired, is a good estimator of  $T_R(\theta_k) - T_F(\theta_k)$ , i.e., of the difference of the marginal IRFs at the *same*  $\theta$ , establishing the usefulness of the interpolation (19).

(24) is called the *regression correction for target ability difference*. Thus, with the correction (24) in place, (13) can be reconstructed, with

$$\hat{\beta}_U = \sum_{k=0}^n \hat{p}_k (\bar{Y}_{Rk}^* - \bar{Y}_{Fk}^*) \quad (25)$$

and  $B$  defined as in (15). Rejection of the hypothesis of no test bias ( $H : \beta_U = 0$ ) occurs when  $B > z_\alpha$ , where  $P[N(0,1) > z_\alpha] = \alpha$  defines  $z_\alpha$ . This procedure will be referred to as the SIB procedure, "SIB" for simultaneous item bias.

Thus, the contribution to the differences  $\bar{Y}_{Rk} - \bar{Y}_{Fk}$  due to target ability difference has been eliminated. It is extremely instructive to note that the correction (24) is the sample analogue of (23), which is basically the decomposition (19), albeit on a different latent scale (though the two latent scales,  $\theta$  and  $V$ , are indistinguishable up to a monotonic transformation).

### A modification of the basic procedure to achieve better statistical behavior

Redefine  $\hat{p}_k$  to be the proportion of all examinees (focal and reference group) attaining  $X = k$ . That is  $\hat{p}_k = (J_{Fk} + J_{Rk}) / \sum_{k=0}^n (J_{Fk} + J_{Rk})$ . Substitute this new  $\hat{p}_k$  into (25) and (16) to obtain the statistic  $B$  of (15). Because of a slightly better adherence in simulation studies to the nominal level of significance when the hypothesis of no test bias holds, this new choice of  $\hat{p}_k$  is recommended over the slightly more intuitive choice based upon focal group examinees alone. The power performance of both versions of  $B$  when test bias was present was very similar. It is upon this version of the SIB statistic that our simulation studies reported below are based.

### SIMULATION STUDY

In order to assess the performance of the procedure in a variety of testing situations, a moderate-sized (84 simulation cases) simulation study was performed. Three parameter logistic item parameters actually estimated from two test data sets, an ACT math test (estimated by Drasgow (1987)) and an ASVAB auto shop test (estimated by Mislevy and Bock (1984)), are used to specify the IRFs in the IRT model. Univariate and bivariate

normal ability distributions, appropriately centered relative to the test item parameters (for the purpose of good measurability of target ability), are used for the focal and reference groups. Two levels of bias and three levels of target ability difference are simulated; tests with a singly-based item and with three biased items are used in the simulations. The level of guessing in the tests is varied. Finally, group size pairs of (3000, 3000), (3000, 1000), and (1500, 1500) for the reference group and focal group examinees respectively are used.

Each simulation model is run 100 times (trials). For a particular simulation model, the item parameters and the two ability distributions for the two groups are fixed; however, at each trial, a new set of examinees (ability parameters) is generated from the ability distributions.

When a single item is to be studied in a simulation, the Mantel-Haenszel procedure as modified by Holland and Thayer is run in parallel in order to provide an external reference to compare to and to compare our procedure with.

### Item parameters

Estimated item parameters from the above mentioned tests were used to construct test models; the ASVAB test length is 25, and the ACT test length is 40. Table 1 gives the summary statistics for the a's, b's, and c's as estimated by Mislevy and Bock and by Drasgow; for the actual parameter values, see Mislevy and Bock (1984) and Drasgow (1987).

place table 1 here

The test for each simulation was generated in the following manner. Let  $N$  denote test length and  $n_b$  the number of items to be studied for possible bias. First,  $n_b$  was chosen to be either 1 or 3. There were two cases to consider.

1. No bias: unidimensional items are used for the entire test.
2. Bias: unidimensional items are used in the valid subtest, and 2-dimensional items are used in the studied subtest.

place table 2 about here

In the first case,  $n_b$  of the  $N$  items were chosen randomly to be the studied ones, and the remainder were used as the valid subtest. In the second case,  $n = N - n_b$  items were chosen at random from either the ASVAB or the ACT test to be the valid subtest, and the 2-dimensional studied item parameters were chosen according to Table 2. Note that the studied item guessing parameters are a function of the average and standard deviation of the guessing parameters on the ASVAB or ACT tests; the studied item  $a$ 's and  $b$ 's are the same for both tests.

The IRFs are for case 1 (no bias)

$$P_i(\theta) = c_i + \frac{(1 - c_i)}{1 + \exp(-1.7a_{i\theta}(\theta - b_{i\theta}))} \quad i = 1, \dots, N, \quad (26)$$

where  $a_{i\theta}$  and  $b_{i\theta}$  are the target discrimination and difficulty for item  $i$ . In case 2 (bias), items 1 to  $n$  were of the form (26), and items  $n + 1$  to  $N$  (studied items) had IRFs

$$P_i(\theta, \eta) = c_i + \frac{(1 - c_i)}{1 + \exp(-1.7(a_{i\theta}(\theta - b_{i\theta}) + a_{i\eta}(\theta - b_{i\eta})))} \quad i = n + 1, \dots, N. \quad (27)$$

The final factor in determining the item parameters was whether or not to include guessing; that is, whether to assume 2PL or 3PL modeling. The presence of guessing is thought to influence the performance of the procedure. Thus, in some simulation models, the estimated  $c_i$ 's from the literature were used in conjunction with (26) and (27); in others all  $c_i$ 's were set to 0 producing a 2PL model. A detailed description of the experimental design of the simulations follows.

### Ability distributions

Specifying the ability distributions involves choosing the five parameters determining the bivariate normal distributions for each group in such a way to meet the following goals:

1. Introduce a specified amount of group difference between target ability distributions.
2. Require the test to measure the target ability well, as would be true for any "good" test.
3. Introduce a specified amount of potential for bias into the distributions.
4. In the case of 2-dimensional studied items (bias case), require that examinee nuisance abilities be influential in determining the response to the item, e.g., that target and reference group examinees have moderate nuisance abilities.

Each goal is elaborated upon separately below. The bivariate distributions for group  $g$  ( $g = R$  or  $F$ ) is denoted

$$\begin{pmatrix} \Theta | g \\ \eta | g \end{pmatrix} \sim N \left[ \begin{pmatrix} \mu_{\theta g} \\ \mu_{\eta g} \end{pmatrix}, \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix} \right] \quad (28)$$

where  $\rho = \text{Corr}(\Theta, \eta | G = g)$  is taken to be the same for both groups ( $\rho$  taken to be different across group tends to introduce bidirectional bias, where marginal IRFs in  $\theta$  for the two groups cross; see Shealy (1989)). Note that  $\sigma^2(\Theta | g)$  and  $\sigma^2(\eta | g)$  are taken to be 1 in our study.

**Goal 1.** We first define target ability difference. We need some notation; let  $\alpha_R =$  the proportion of the entire (conceptual) population of examinees who are reference group members, and  $\alpha_F = 1 - \alpha_R$  be the corresponding proportion for the focal group. (Note: as  $J_R$  and  $J_F$  both increase to  $\infty$ , conceptually,  $\frac{J_R}{J_R + J_F} \rightarrow \alpha_R$  and  $\frac{J_F}{J_R + J_F} \rightarrow \alpha_F$ . Here  $J_g$  denotes the number of sampled Group  $g$  examinees.) Define

$$d_T = \frac{\mu_{\theta R} - \mu_{\theta F}}{\sigma_{\theta P}} \quad (29)$$

to be the *target ability difference* between the focal and reference groups, where

$$\sigma_{\theta P}^2 = \alpha_R \sigma^2(\Theta | R) + \alpha_F \sigma^2(\Theta | F). \quad (30)$$

Note that when (28) holds  $\sigma_{\theta P}^2 = 1$  and thus that  $d_T = \mu_{\theta R} - \mu_{\theta F}$ .  $d_T$  is a quantity specified in the simulations.

**Goal 2.** The criterion used to ensure good measurability of  $\theta$  by the test, is that the average difficulty ( $\bar{b}$ ) of the valid subtest should be close to the average target ability over the pooled groups. Specifically,  $\mu_{\theta R}$  and  $\mu_{\theta F}$  are chosen so that

$$\bar{b} = E[\Theta] \equiv \alpha_R \mu_{\theta R} + \alpha_F \mu_{\theta F}. \quad (31)$$

$\bar{b}$  is taken from Table 1.  $\mu_{\theta R}$  and  $\mu_{\theta F}$  are completely determined by specification of  $d_T$  and (31).

**Goal 3.** We use a more restrictive version of Definition 1 to define potential for bias: set

$$C_\beta(\theta) = E[\eta | \Theta = \theta, G = R] - E[\eta | \Theta = \theta, G = F]. \quad (32)$$

$C_\beta(\theta) > 0$  is defined to be the potential for bias against the focal group. When (28) holds, (32) becomes

$$\begin{aligned} C_\beta(\theta) &\equiv C_\beta = \mu_{\eta R} - \rho \mu_{\theta R} - (\mu_{\eta F} - \rho \mu_{\theta F}) \\ &= (\mu_{\eta R} - \mu_{\eta F}) - \rho(\mu_{\theta R} - \mu_{\theta F}) = (\mu_{\eta R} - \mu_{\eta F}) - \rho d_T, \end{aligned} \quad (33)$$

$\theta$  dropping out because the ability correlation ( $\rho$ ) is equal for both groups. Note that because  $C_\beta$  is constant for all  $\theta$ , unidirectional bias is being introduced. For a specified amount of  $C_\beta$ ,  $\mu_{\eta R}$  and  $\mu_{\eta F}$  are determined partially. The reader should note that potential for bias can hold even though  $\mu_{\eta R} = \mu_{\eta F}$  unless  $\mu_{\theta F} = \mu_{\theta R}$ .

**Goal 4.** The criterion used to ensure nuisance determinant influence is the following. The nuisance difficulties for all studied items were chosen to be 0. For an arbitrarily chosen target ability (say  $\theta = 0$ ) we thus want the average nuisance ability to be near 0 as well. Thus we choose

$$E[\eta | \Theta = 0, G = R] = -E[\eta | \Theta = 0, G = F] \quad (34)$$

i.e., the conditional nuisance expectation at  $\Theta = 0$  is to be centered around the average studied item nuisance difficulty of 0, for the reference and focal groups. Our intent in this study was to introduce bias against the focal group, so  $E[\eta | \theta, R] > 0$  in (34) and thus we get

$$0 < \mu_{\eta R} - \rho\mu_{\theta R} = -(\mu_{\eta F} - \rho\mu_{\theta F}); \quad (35)$$

this will specify  $\mu_{\eta R}$  and  $\mu_{\eta F}$ , along with specification of  $C_\beta$  in (33).

There is an additional issue here: how large should  $C_\beta$  be chosen to introduce a "moderate" or "severe" amount of bias into the 2-dimensional studied items of Table 2? This is treated below, in the experimental design of the study.

Goals 1-4 now completely specify (28):  $\mu_{\theta R}$ ,  $\mu_{\theta F}$ ,  $\mu_{\eta R}$ , and  $\mu_{\eta F}$  can be found by solving (29), (31), (33), and (35) simultaneously for them.  $\rho$ ,  $\sigma^2(\theta | g)$ , and  $\sigma^2(\eta | g)$  are chosen:  $\rho = .5$ , and all  $\sigma$ 's are 1.

### Choice of $C_\beta$

The amount of potential for bias  $C_\beta$  in each simulation model was chosen so that the actual level of bias  $\beta_U$  produced was such that the power behavior of the statistic can be well assessed for the given examinee sample sizes, valid subtest used (recall Table 1), and biased items used (recall Table 2). These  $\beta_U$  values (rounded to two significant figures) are shown in Table 3. The governing equations determining  $C_\beta$  from  $\beta_U$  were

$$\beta_U = \int_{\theta} (T_R(\theta) - T_F(\theta)) f_F(\theta) d\theta$$

where

$$T_g(\theta) = \sum_{i=n+1}^N E[P_i(\Theta, \eta) | \Theta = \theta, G = g] \quad (36)$$

with  $P_i(\theta, \eta)$  defined in (27) and the item parameters in (27) defined in Table 2, and the

place table 3 about here

parameters of the  $(\Theta, \eta)$  distribution determined from (29), (31), (33), and (35). One standard often used to interpret from a practitioner's viewpoint the magnitude of the bias is that the bias is "moderate" if  $0.5 \leq \Delta_{MH} < 1$  while it is "large" if  $\Delta_{MH} \geq 1$ , where  $\Delta_{MH}$  is the theoretical index based on use of the Mantel-Haenszel log odds ratio proposed by Holland and Thayer (1988). The rationale for  $\Delta_{MH}$  and  $\beta_U$  are different, but for  $n_b = 1$  and unidirectional bias, they tend to be highly correlated and are crudely related by

$$\beta_U \doteq \Delta_{MH}/10.$$

Thus, roughly,  $0.05 \leq \beta_U < 0.1$  would constitute moderate bias while  $\beta_U \geq 0.1$  would constitute large bias. Thus in the  $n_b = 1$  case, referring to Table 4, the amount of bias being simulated is actually either (low) moderate or small. Examination of (36) shows that  $\beta_U$  is a measure of how much lower the probability of getting the biased item right is for an average focal group examinee as compared with an average reference group examinee of the same target ability. Thus  $\beta_U$  has a natural and useful empirical interpretation. In our context,  $\Delta_{MH}$ , by contrast, is a measure of horizontal distance between  $T_R(\theta)$  and  $T_F(\theta)$  at  $y = \frac{1+\bar{c}}{2}$  (i.e., the value of  $T_R^{-1}((1 + \bar{c})/2) - T_F^{-1}((1 + \bar{c})/2)$ ), where  $\bar{c}$  is defined in Table 1.

place table 4 about here

### Experimental design

The design is as follows. For the case of no test bias ( $C_\beta = 0$ ), for each test type

(ASVAB Auto Shop or ACT Math) the following simulations are done:

$$n_b = \left\{ \begin{array}{c} 1 \\ 3 \end{array} \right\} \times d_T = \left\{ \begin{array}{c} 0.0 \\ 0.5 \\ 1.0 \end{array} \right\} \times J_R/J_F = \left\{ \begin{array}{c} 3000/3000 \\ 3000/1000 \\ 1500/1500 \end{array} \right\}$$

$$\supset \left\{ \begin{array}{c} \text{guessing} \\ \text{no guessing} \end{array} \right\}.$$

Here “guessing” means that the estimated ACT and ASVAB guessing parameters are used in the model and “no guessing” means that all  $c_s$  are set to zero; that is, 2PL modeling is used. Also, “ $\supset$ ” means that this guessing “factor” is randomly assigned within the 36 levels produced by crossing the other factors.

For the case of test bias ( $C_\beta > 0$ ) the following simulation are done for each test type:

$$n_b = \left\{ \begin{array}{c} 1 \\ 3 \end{array} \right\} \times d_T = \left\{ \begin{array}{c} 0.0 \\ 0.5 \end{array} \right\} \times C_\beta = \left\{ \begin{array}{c} 0.5 \\ 1.0 \end{array} \right\} \times J_R/J_F = \left\{ \begin{array}{c} 3000/3000 \\ 3000/1000 \\ 1500/1500 \end{array} \right\}$$

$$\supset \left\{ \begin{array}{c} \text{guessing} \\ \text{no guessing} \end{array} \right\}.$$

For  $n_b = 1$ , the nuisance discrimination  $a_{N\eta}$  of the studied item is .8; for  $n_b = 3$ , the nuisance discrimination of each of the 3 studied items is .4. These discriminations were chosen so that the power of the procedure could be well assessed (i.e., so that it would not be *too* close to 1). It is informative to note in passing that the power of the procedure is expected to be greater when  $n_b$  is increased from 1 to 3 unless each item individually displays less bias in the  $n_b = 3$  case. This is why the  $a_{i\eta}$  ( $i = N - 2, N - 1, N$ ) was chosen to be .4 in the  $n_b = 3$  case,  $\frac{1}{2}$  of that used in the  $n_b = 1$  case.

There are therefore 48 simulation models that incorporate bias. Thus, a total of 84 simulation models were used in the simulation study.

## RESULTS OF THE SIMULATION STUDY

The results of the simulation study are given in Tables 5–8 and 9–12, with Tables 5–8 summarizing the no test bias simulations and Tables 9–12 summarizing the simulations having test bias present. The  $c$  column indicates whether the model has guessing present or not. In all  $n_b = 1$  cases, the Mantel-Haenszel rejection rate for the hypothesis of no item bias (based on 100 trials) is reported in the MH column. In all cases the SIB rejection rate is reported in the SIB column. In all cases where test bias is present (Tables 9–12), the  $C_\beta$  column presents the amount of potential for bias present (recall (33)); the  $\beta_U$  column presents our index of the amount of bias present against the focal group in the model

(recall (6));  $\bar{\hat{\beta}}_U$  is the average of the estimates  $\hat{\beta}_U$  of  $\beta_U$  over the 100 trials; the  $\Delta_{MH}$  column presents the amount of bias present against the focal group in the model from the Mantel-Haenszel perspective.

Tables 5–8 indicate that both the SIB statistic and the MH statistic display reasonable adherence to the nominal level of significance of 0.05. There appear to be situations of no bias, which have a target ability difference and which depart from the Rasch model, where the Mantel-Haenszel procedure displays inflated Type 1 error. (See Zwick (1990), for a discussion of this problem and an illustrative example.) There is evidence that in such situations (Shealy (1989)), the SIB statistic adheres closely to the nominal level of significance. On the other hand there are likely portions of the “parameter space” of realistic IRT models where our linear regression correction is stressed and hence the MH would likely display better Type 1 error performance. More study is required before it can be claimed that either MH or SIB displays superior Type 1 error performance. The striking fact is that *both* procedures seem to be quite robust against the inflating Type 1 error effect of differing target ability distributions. In this regard,  $d_T = 1$  from the practitioner’s perspective is certainly a large amount of target ability difference.

Tables 9 and 11 indicate that both the SIB statistic and the MH statistic are quite powerful against moderate amounts of bias and fairly powerful against small amounts of bias in a single biased item. Untabulated simulation studies for larger amounts of bias produced rejection rates of essentially unity for both the SIB and MH procedures.

Tables 10 and 12 indicate that the SIB procedure is quite powerful against moderate amounts of bias resulting from several (3 here) items producing bias in the same direction. The reader should recall that the amount of bias/item was lowered for the  $n_b = 3$  case by reducing the discrimination in the nuisance dimension from  $a_{\eta_N} = 0.8$  to  $a_{\eta_i} = 0.4$  for the studied items. In both the  $n_b = 1$  and  $n_b = 3$  cases, the potential for bias as measured by  $C_\beta$  was kept the same ( $C_\beta = 0.2$  or  $0.3$ ). These two tables show, as claimed, that the SIB procedure can successfully detect simultaneous item bias, even if the amount of bias present per item is small.

Tables 9 and 11 show, for the particular bias models of the simulation study, that SIB is somewhat more powerful than MH, averaging 0.07 higher for those models for which rejection rates are  $< 0.9$ . We do not know whether this greater SIB power generalizes to other models of bias.

Tables 9–12 provide evidence about the ability of  $\hat{\beta}_U$  to estimate  $\beta_U$ , our measure of the amount of bias present. For each case  $\bar{\hat{\beta}}_U$  is an indicator of the amount of *statistical bias* one might expect in using  $\hat{\beta}_U$ . Clearly statistical bias of roughly  $+0.01$  is present. The estimated standard errors for  $\hat{\beta}_U$  are not recorded, but averaged (roughly) about  $1/3$  of  $\bar{\hat{\beta}}_U$ . Thus if  $\hat{\beta}_U = 0.05$  there is likely a bias of 0.01 and a standard error of 0.017. Thus, crudely, a 95% confidence interval (if asymptotic normality is a good approximation) would

be given by  $0.04 \pm 0.028$ . Here  $0.04 = 0.05 - 0.01$  is the correction for statistical bias. It would seem that  $\hat{\beta}_U$  provides a useful empirical index of the amount of bias present in a statistical subtest of items; more work is planned in studying its theoretical and empirical properties.

## SUMMARY AND CONCLUSIONS

The SIB procedure was designed to test for unidirectional test bias residing in one or more items, using the conception that test bias is incipient within the two groups' ability distributions (in terms of a difference in conditional nuisance ability distributions). By means of the regression correction presented here, the inflation of the SIB test statistic due to target ability difference (one group having a stochastically larger distribution of  $\Theta$ ) is extracted. This correction represents a conceptual link between conditional-on-observed-score methods and IRT-based methods, just as the practice of including the studied item in the comparable examinee criterion in the Mantel-Haenszel procedure of Holland and Thayer (1988) does. The correction adjusts the studied subtest scores for the two groups so that they are now estimates of the *same* latent IRT ability in the case of no test bias, even if group target abilities exist. It is useful to note that the adjustment, although conceptually based upon multidimensional IRT modeling, is in fact computed using a classical approach and hence does not depend on IRT ability or item parameter estimation.

A moderate (84 models) simulation study shows that both MH and SIB display good adherence to the nominal level of significance, even for large ( $d_T = 1$ ) target ability differences. In the case of a single biased item, both MH and SIB display good power with SIB displaying slightly higher power. As designed, the SIB statistic displays good power in the case of several biased items (3 here), even when the amount of bias/item is fairly small.

A large scale simulation study is in progress with the goal of obtaining a better understanding of the performance characteristics of both the SIB and the MH statistics with particular emphasis on investigation of statistical power and adherence to the nominal level of significance. Based upon the completed portion of this simulation study reported herein, we would recommend that practitioners use the SIB and MH statistics simultaneously. Both are extremely easy to compute and for moderate sized data sets run quickly on a typical PC configuration. Carefully checked code with a user oriented driver is available from the authors for running both the SIB and MH statistics on real data sets and also for doing simulation studies of performance.

## APPENDIX

1. Derivation of  $\hat{V}_g(k)$ , the estimated regression of true on observed valid subtest score, for  $k = 0, \dots, n$ .

Recall that  $V_g(k) = E[\bar{P}(\Theta) | k, g]$  needs to be estimated in order for  $S_g(V_k)$  of (23) to be estimated. Suppressing  $g$  for simplicity, we need to estimate  $V(k)$  at  $k = 0, 1, \dots, n$ . Although  $V(k)$  is not necessarily linear in  $k$  (see Shealy (1989), p. 87ff for a discussion), as an approximation we assume  $nV(k)$  is linear in  $k$ ; i.e.,

$$nV(k) = \alpha + \beta k.$$

To estimate  $V(k)$ , we consider the true score model for the valid subtest score  $X$ :

$$X = T + e \quad (\text{A1})$$

where

$$E(e) = 0, \quad \text{cov}(T, e) = 0 \quad (\text{A2})$$

is assumed and the true score  $T$  has the latent variable representation  $T = n\bar{P}(\Theta)$ . Thus

$$nV(k) = E[T | k].$$

Standard regression theory for  $E(T | k)$  yields

$$V(k) = \frac{1}{n} \left( ET + \frac{\rho_{XT}\sigma_T}{\sigma_X} (k - EX) \right). \quad (\text{A3})$$

But, for the true score model given by (A1) and (A2),

$$\frac{\rho_{XT}\sigma_T}{\sigma_X} = 1 - \frac{\sigma^2(e)}{\sigma^2(X)}, \quad (\text{A4})$$

is well known (see page 61 of Lord and Novick (1968)). Using (A1) and (A2),  $ET = EX$  holds. Thus, by (A3) and (A4),

$$V(k) = \frac{1}{n} \left[ EX + \left( 1 - \frac{\sigma^2(e)}{\sigma^2(X)} \right) (k - EX) \right] \quad (\text{A5})$$

holds.

Clearly  $EX \equiv E[X | g]$  can be estimated by the average valid subtest score  $\bar{X}_g$  of all Group  $g$  examinees taking the test. Thus it remains to estimate  $\sigma^2(e)/\sigma^2(X)$ .

$\sigma^2(X) \equiv \sigma^2(X | g)$  can clearly be estimated by the usual sample variance estimate of all Group  $g$  examinees taking the test

$$\hat{\sigma}^2(X | g) \stackrel{\text{def}}{=} \frac{1}{(J_g - 1)} \sum_{j=1}^{J_g} (X_{gj} - \bar{X}_g)^2, \quad (\text{A6})$$

where  $J_g$  denotes the number of Group  $g$  examinees taking the test and  $X_{gj}$  is the valid subtest number correct score of the  $j$ th such Group  $g$  examinee. It remains to estimate  $\sigma^2(e)$ ; denote this estimation by  $\hat{\sigma}^2(e)$ . Then the desired estimation of  $\sigma^2(e)/\sigma^2(X)$  will be given by  $\hat{\sigma}^2(e)/\hat{\sigma}^2(X)$ . A standard conditioning formula yields, indexing the valid subtest items by  $i = 1, 2, \dots, n$ , and setting  $X_g = X | g$ ,  $\Theta_g = \Theta | g$  as a reminder that sampling here is from Group  $g$  only,

$$\begin{aligned} \sigma^2(X | g) \equiv \sigma^2(X_g) &= \sigma^2(E[X_g | \Theta_g]) + E[\sigma^2(X_g | \Theta_g)] \\ &= \sigma^2(n\bar{P}(\Theta_g)) + \sum_{i=1}^n E[P_i(\Theta_g)(1 - P_i(\Theta_g))], \end{aligned} \quad (\text{A7})$$

using the standard item response theory assumption of local independence of items, given  $\Theta$ . Also, by (A2) it is trivial that

$$\sigma^2(X | g) = \sigma^2(n\bar{P}(\Theta) | g) + \sigma^2(e | g).$$

Thus, by (A7),

$$\sigma^2(e | g) = \sum_{i=1}^n E[P_i(\Theta_g)(1 - P_i(\Theta_g))].$$

This suggests

$$\hat{\sigma}^2(e | g) = \sum_{i=1}^n \bar{U}_{ig}(1 - \bar{U}_{ig}), \quad (\text{A8})$$

where  $\bar{U}_{ig}$  is the proportion correct for Group  $g$  examinees for valid subtest item  $i$ . Thus, using (A5), we will estimate  $V_g(k)$  by

$$\hat{V}_g(k) = \frac{1}{n} \left[ \bar{X}_g + \left( 1 - \frac{\hat{\sigma}^2(e | g)}{\hat{\sigma}^2(X | g)} \right) (k - \bar{X}_g) \right]. \quad (\text{A9})$$

**2. The complete procedure to detect test bias, using the proposed regression correction.**

The SIB procedure in its entirety is presented here. First we set some basic notation. Group  $g$  ( $g = R$  or  $F$ ) has  $J_g$  examinees taking the test of  $N$  items. The response to item  $i$  of the  $j$ th group  $g$  examinee is  $U_{gij}$ . The subtest scores are

$$X_{gj} = \sum_{i=1}^n U_{gij} \quad (\text{valid subtest score}), \quad Y_{gj} = \sum_{i=n+1}^N U_{gij} \quad (\text{studied subtest score}).$$

The classical group item difficulties are  $\bar{U}_{gi} = (1/J_g) \sum_{j=1}^{J_g} U_{gij}$ . Let  $\sum_j^{(k)}$  denote summation over those group  $g$  examinees  $j$  with  $k$  correct on the valid subtest.

1. Compute  $J_{gk}$ , the number of group  $g$  examinees with  $k$  correct on the valid subtest.
2. Compute

$$\bar{Y}_{gk} = \frac{1}{J_{gk}} \sum_j^{(k)} Y_{gj}$$

$$S_{gk}^2 = \frac{1}{J_{gk} - 1} \sum_j^{(k)} (Y_{gj} - \bar{Y}_{gk})^2.$$

- If  $J_{gk} = 0$ , set  $\bar{Y}_{gk} = 0$ ; if  $J_{gk} \leq 1$ , set  $S_{gk}^2 = 0$ .  $\bar{Y}_{gk}$  is the sample average studied subtest score of group  $g$  examinees attaining  $X_g = k$ , and  $S_{gk}^2$  is the sample variance.
3. Compute  $\hat{P}_g(k) = J_{gk}/J_g$ , for both groups and all  $k$ .  $\hat{P}_g(k)$  is the estimate of the histogram of  $X | G = g$ . Then compute  $\hat{P}_g^*(k)$ , the MLE of the unimodalized histogram of  $X | G = g$ , over the class of all possible unimodal MLE of the histograms with  $n + 1$  possible values ( $X | G = g$  is assumed to have a unimodal distribution and hence its estimate  $\{\hat{P}_g^*(k), k \geq 0\}$  should also be unimodal). For details of this procedure, using the up-and-down-blocks algorithm, see Barlow et al. (1972; pp. 72-73; pp. 223-231).
  4. Set  $I(k) = 1$  for all  $k$  unless either
    - (a)  $k = 0$  or  $n$ ,
    - (b)  $S_{Rk}^2 = 0$  or  $S_{Fk}^2 = 0$ ,
    - (c)  $J_R \hat{P}_R^*(k) < J_{\min}$  or  $J_F \hat{P}_F^*(k) < J_{\min}$  where  $J_{\min}$  is set by user, usually around 30, or
    - (d)  $k \leq nc_U$ , where  $c_U \geq 0$  is the user-specified global guessing parameter for the test. (It is assumed that there is a relatively constant level of guessing across item, and that there is at least partial knowledge of this guessing value.)

$I(k)$ ,  $k = 0, \dots, n$ , is the *examinee inclusion* indicator; it is 1 if examinees with  $X = k$  are to have their responses included in the test statistic. (a) excludes the two extreme valid subtest scores because of their poor estimation of target ability. The (b) exclusion is obvious. The (c) exclusion is done to assure that each valid subtest score category has enough examinees to make  $\bar{Y}_{Rk}$  and  $\bar{Y}_{Fk}$  approximately normal; the unimodal mass function is used so that only extreme valid subtest score categories are excluded. As for (d), all valid scores below that expected by guessing are excluded.

5. Compute the regression of true score on valid subtest score:
  - (a)  $\bar{U}_{gi}^* = \frac{\bar{U}_{gi} - c_U}{1 - c_U}$ . If the result is  $< 0$ , set it to 0 (adjustment for guessing).
  - (b)  $\bar{X}_g = \frac{1}{J_g} \sum_{j=1}^{J_g} X_{gj}$
  - (c)  $\hat{\sigma}^2(X | g) = \frac{1}{J_g - 1} \sum_{j=1}^{J_g} (X_{gj} - \bar{X}_g)^2$
  - (d)  $\hat{\sigma}^2(e | g) = \sum_{i=1}^n \bar{U}_{gi}^* (1 - \bar{U}_{gi}^*)$
  - (e)  $\hat{b}_g = \frac{n}{n-1} \left( 1 - \frac{\hat{\sigma}^2(e|g)}{\hat{\sigma}^2(X|g)} \right)$

(f)  $\hat{V}_g(k) = \frac{1}{n}(\bar{X}_g + \hat{b}_g(k - \bar{X}_g))$  for both  $g$  and  $k = 0, \dots, n$ .

6. Make the regression correction:

(a)  $k_\ell = \min\{k : I(k) = 1\}$ ,  $k_r = \max\{k : I(k) = 1\}$ .

(b)  $\hat{V}_k = \frac{1}{2}(\hat{V}_R(k) + \hat{V}_F(k))$ , for  $k_\ell \leq k \leq k_r$ .

(c) For  $k_\ell < k < k_r$ , compute

$$\hat{M}_{gk} = \frac{\bar{Y}_{g,k+1} - \bar{Y}_{g,k-1}}{\hat{V}_g(k+1) - \hat{V}_g(k-1)}.$$

Then compute  $\bar{Y}_{gk}^* = \bar{Y}_{gk} + \hat{M}_{gk}(\hat{V}_k - \hat{V}_g(k))$ .

(d) For  $k = k_\ell$  and  $k = k_r$ , compute  $\bar{Y}_{gk}^*$  in the following way.

i. Define

$$\hat{S}_g(v) = \begin{cases} (1 - \alpha)\bar{Y}_{g,k+1} + \alpha\bar{Y}_{gk} & \text{if } \hat{V}_g(k) \leq v < \hat{V}_g(k+1) \\ \bar{Y}_{g0} & \text{if } v < \hat{V}_g(0) \\ \bar{Y}_{gn} & \text{if } v \geq \hat{V}_g(n), \end{cases}$$

and

$$\alpha = \frac{v - \hat{V}_g(k)}{\hat{V}_g(k+1) - \hat{V}_g(k)}.$$

$\hat{S}_g(v)$  is the linear interpolation of  $\{\bar{Y}_{g0}, \dots, \bar{Y}_{gn}\}$ .

ii. Compute

$$\bar{Y}_{gk}^* = \hat{S}_g(\hat{V}_k)$$

for  $k = k_\ell$  and  $k = k_r$ .

7. Compute the bias statistic.

(a) Compute  $J_g^* = \sum_{k=0}^n I(k)J_{gk}$ , the number of included group  $g$  examinees

(b) Compute

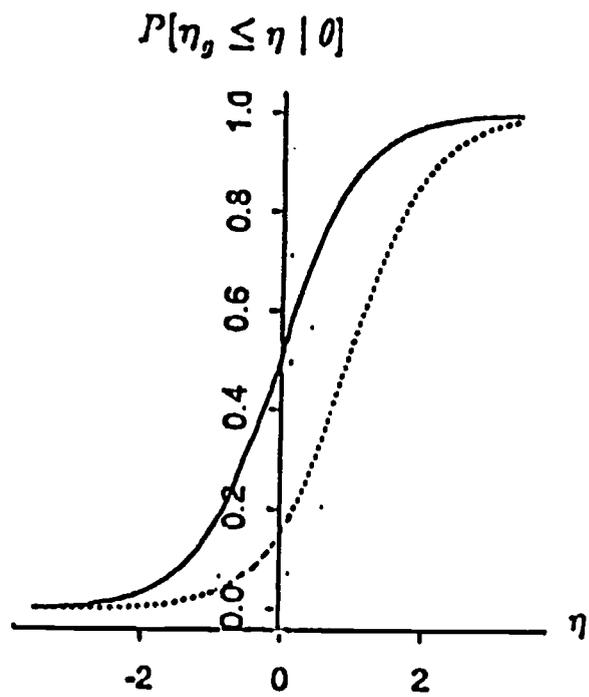
$$B = \frac{\sum_{k=0}^n \frac{J_{Fk}}{J_F^*} (\bar{Y}_{Rk}^* - \bar{Y}_{Fk}^*) I(k)}{\left( \sum_{k=0}^n \frac{J_{Fk}^2}{J_F^{*2}} (S_{Rk}^2 + S_{Fk}^2) I(k) \right)^{1/2}}.$$

(c) Reject  $H : \beta_U = 0$  in favor of  $\beta_U > 0$  at level  $\alpha$  if  $B > z_\alpha$ , where  $P[N(0,1) > z_\alpha] = \alpha$  defines  $z_\alpha$ .

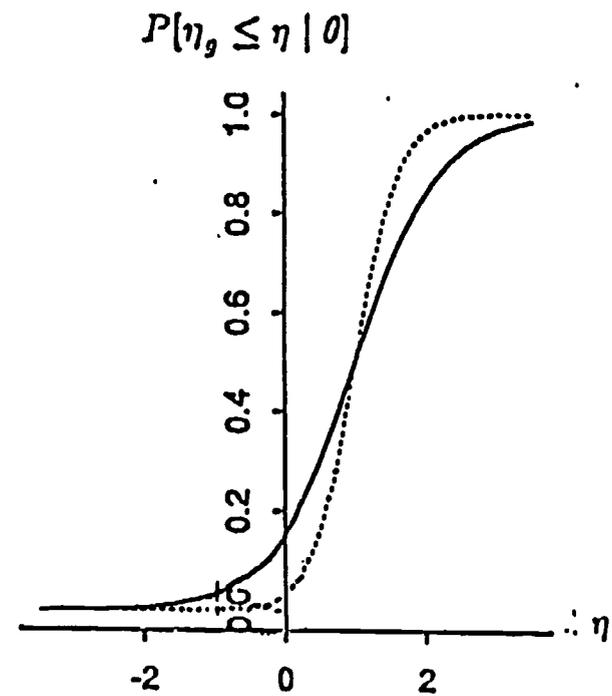
## References

- Ackerman, T. (1991). A didactic explanation of item bias, item impact, and item validity from a multidimensional IRT perspective. Submitted for publication and presented at 1991 annual AERA/NCME joint meeting.
- Ansley, T.N. and Forsyth, R.A. (1985). An examination of the characteristics of unidimensional IRT parameter estimates derived from two-dimensional data. *Applied Psychological Measurement* 9, 37-48.
- Barlow, R., Bartholomew, D., Bremner, J., and Brunk, H. (1972). *Statistical Inference under Order Restrictions*. New York: John Wiley.
- Dragow, F. (1987). A study of measurement bias of two standard psychological tests. *Journal of Applied Psychology* 72, 19-30.
- Hambleton, R.K. and Swaminathan, H. (1985). *Item Response Theory: Principles and Applications*. Boston: Kluwer-Nijhoff Publishing.
- Holland, P.W. and Thayer, D.T. (1988). Differential item functioning and the Mantel-Haenszel procedure. In H. Wainer and H.I. Braun (Eds.), *Test Validity*, (pp. 129-145). Hillsdale, New Jersey: Lawrence Erlbaum.
- Kok, F. (1988). Item Bias and Test Multidimensionality. In R. Langeheine and J. Rost (Eds.), *Latent Trait and Latent Models*, (pp. 263-275). New York: Plenum Press.
- Lautenschlager, G. and Park, D. (1988) IRT item bias detection procedures: issues of model mis-specification, robustness, and parameter linking. *Applied Psychological Measurement* 12, 365-376.
- Linn, R.L. and Harnish, D. (1981). Interactions between item content and group membership on achievement test items. *Journal of Educational Measurement* 18, 109-118.
- Linn, R., Levine, M., Hastings, C., and Wardrop, J. (1981). Item bias on a test of reading comprehension. *Applied Psychological Measurement* 5, 159-173.
- Lord, F.M. (1980). *Applications of Item Response Theory to Practical Testing Problems*. Hillsdale, New Jersey: Lawrence Erlbaum.
- Lord, F.M. and Novick, M.R. (1968). *Statistical Theories of Mental Test Scores*. Reading, Massachusetts: Addison-Wesley.
- Mellenbergh, G.J. (1982). Contingency table methods for assessing item bias. *Journal of Educational Statistics* 7, 105-118.
- Millsap, R.E. and Meredith, W. (1989). *The Detection of DIF: Why There is No Free Lunch*. Paper presented at the Annual Meeting of the Psychometric Society, University of California at Los Angeles, July 6-9, 1989.
- Mislevy, R.J. and Bock, R.D. (1984). Item operating characteristics of the Armed Services Aptitude Battery (ASVAB). Form 8A. Office of Naval Research Technical Report (N00014-83-C-0283).
- Shealy, R.T. (1991). Assessment of the Shealy-Stout test bias statistic: a simulation study. In preparation.

- Shealy, R.T. (1989). *An Item Response Theory-Based Statistical Procedure for Detecting Concurrent Internal Bias in Ability Tests*. Unpublished doctoral dissertation, University of Illinois, Urbana-Champaign.
- Shealy, R.T. and Stout, W.F. (1991). *An Item Response Theory Model for Test Bias* (Technical Report 4421-548 under ONR grant N00014-90-J-1940). Champaign, Urbana: Department of Statistics, University of Illinois (A 1989 version of this was widely distributed; it will appear, by invitation, in *Differential Item Functioning, Theory and Practice*, 1992, Hillsdale, New Jersey: Erlbaum.)
- Thissen, D., Steinberg, L., and Wainer, H. (1988). Use of item response theory in the study of group differences in trace lines. In H. Wainer and H.I. Braun (Eds.), *Test Validity* (pp. 147-169). Hillsdale, New Jersey: Lawrence Erlbaum.
- Zwick, R. (1990). When do item response function and Mantel-Haenszel definitions of differential item functioning coincide? *Journal of Educational Statistics* 15, 185-197.



stochastically ordered



not stochastically ordered

Figure 1. Stochastically ordered and unordered pairs of distributions

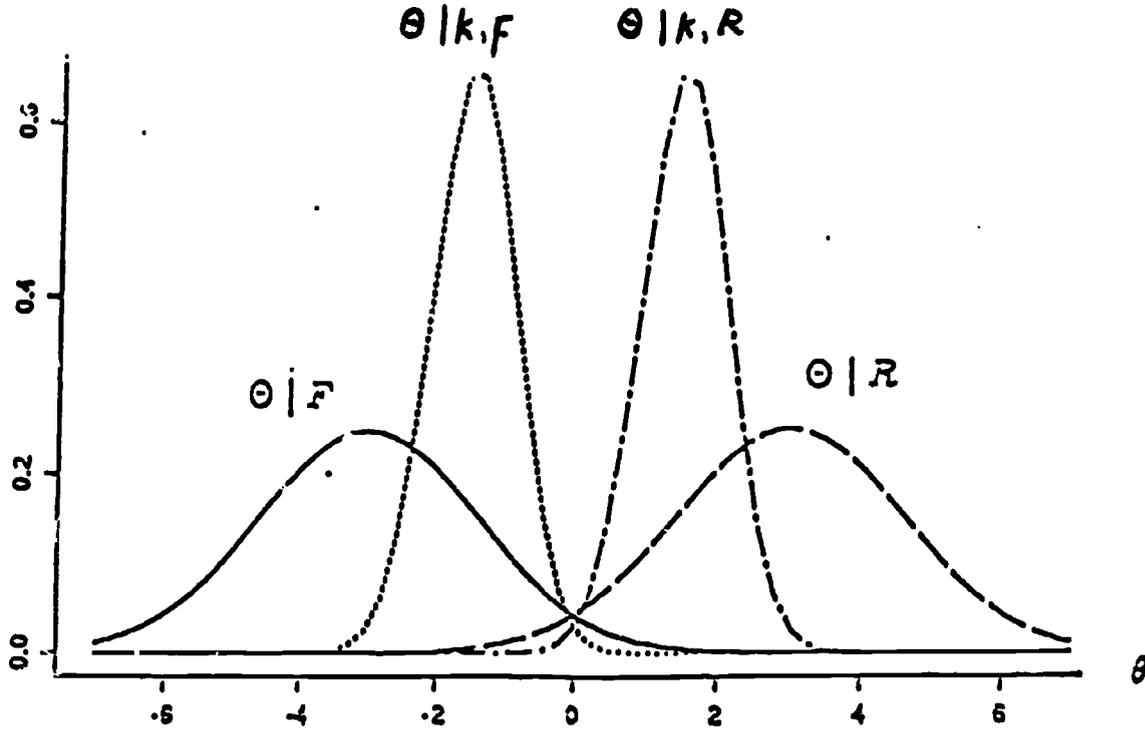


Figure 2. Prior and posterior target ability distributions

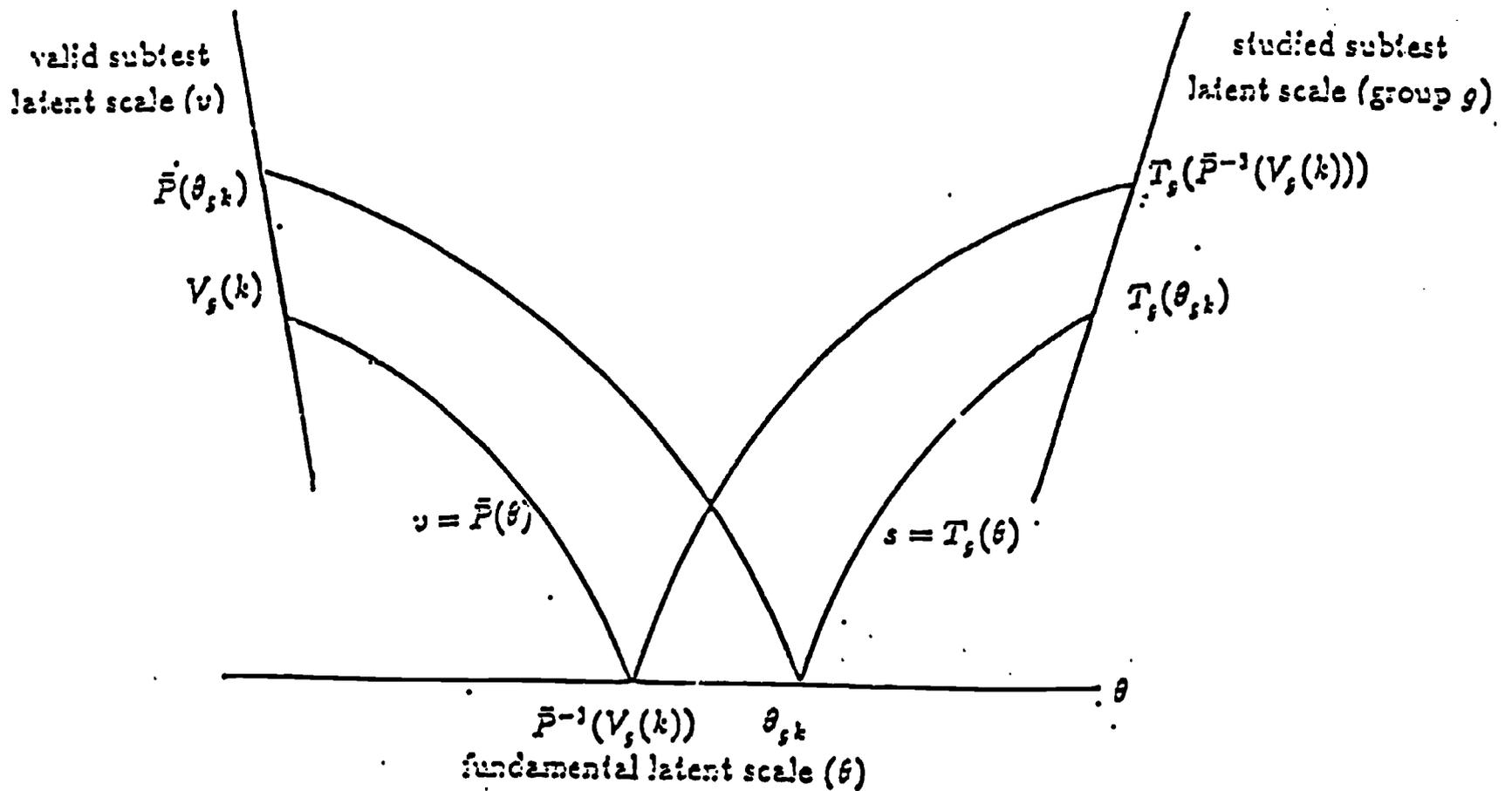


Figure 3. The three latent scales.

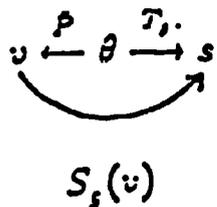


Figure 4. The valid subtest to studied subtest transformation

Table 1: Means and sds for the ASBAB and ACT item parameters used in the study.

Test	$\bar{a}$	$\sigma_a$	$\bar{b}$	$\sigma_b$	$\bar{c}$	$\sigma_c$	N
ASVAB auto/shop	1.22	0.7	0.09	0.72	0.20	0.06	25
ACT math	1.09	0.35	0.5	0.61	0.14	0.04	40

Table 2: Item parameters for 2-dimensional studied in the bias case.

$n_b$	Item No.	$a_{i\theta}$	$b_{i\theta}$	$a_{i\eta}$	$b_{i\eta}$	$c_i$
1	$N$	1.0	0.0	0.8	0.0	$\bar{c}$
3	$N-2$	0.6	-0.3	0.4	0.0	$\bar{c} - \frac{1}{2}\sigma_c$
	$N-1$	0.8	0.0	0.4	0.0	$\bar{c}$
	$N$	1.0	0.3	0.4	0.0	$\bar{c} + \frac{1}{2}\sigma_c$

Table 3: Equivalence table for bias potential and actual test bias.

$n_b$	$C_\beta$	$a_\eta$	$\beta_U$
1	0.0	-	0
1	0.2	0.8	0.03
1	0.3	0.8	0.05
3	0.0	-	0
3	0.2	0.4	0.06
3	0.3	0.4	0.09

Table 4: Equivalence of  $\Delta_{MH}$  and  $\beta_U$  when  $n_b = 1$ , using item parameters of Table 2.

$C_\beta$	$c$ 's used	$\Delta_{MH}$	$\beta_U$
0.0	-	0	0
0.2	0.0	.27	0.034
0.2	actual $c$ 's	.27	0.026
0.3	0.0	.40	0.051
0.3	actual $c$ 's	.39	0.039

Table 5: No bias, ACT,  $n_b = 1$ ,  $\alpha = 0.05$ .

$J_F$	$J_R$	$c$	$d_T$	MH	SIB
1500	1500	0	.0	.03	.07
1000	3000	0	.0	.00	.02
3000	3000	$c$	.0	.09	.06
1500	1500	0	.5	.04	.04
1000	3000	$c$	.5	.10	.10
3000	3000	$c$	.5	.05	.03
1500	1500	$c$	1.0	.02	.05
1000	3000	$c$	1.0	.05	.10
3000	3000	0	1.0	.06	.09

Table 6: No bias, ACT,  $n_b = 3$ ,  $\alpha = 0.05$ .

$J_F$	$J_R$	$c$	$d_T$	SIB
1500	1500	0	.0	.05
1000	3000	0	.0	.02
3000	3000	$c$	.0	.07
1500	1500	0	.5	.08
1000	3000	$c$	.5	.07
3000	3000	0	.5	.05
1500	1500	$c$	1.0	.06
1000	3000	$c$	1.0	.16
3000	3000	0	1.0	.09

Table 7: No bias, ASVAB,  $n_b = 1$ ,  $\alpha = 0.05$ .

$J_F$	$J_R$	$c$	$d_T$	MH	SIB
1500	1500	0	.0	.08	.07
1000	3000	0	.0	.04	.04
3000	3000	$c$	.0	.06	.06
1500	1500	0	.5	.13	.14
1000	3000	$c$	.5	.04	.03
3000	3000	$c$	.5	.05	.04
1500	1500	$c$	1.0	.07	.02
1000	3000	$c$	1.0	.15	.09
3000	3000	0	1.0	.11	.01

Table 8: No bias, ASVAB,  $n_b = 3$ ,  $\alpha = 0.05$ .

$J_F$	$J_R$	$c$	$d_t$	SIB
1500	1500	0	.0	.07
1000	3000	0	.0	.04
3000	3000	$c$	.0	.03
1500	1500	0	.5	.07
1000	3000	$c$	.5	.06
3000	3000	0	.5	.05
1500	1500	$c$	1.0	.15
1000	3000	$c$	1.0	.07
3000	3000	0	1.0	.04

Table 9: Bias,  $a_7 = 0.8$ , ACT,  $n_b = 1$ ,  $\alpha = 0.05$ .

$J_F$	$J_R$	$c$	$d_T$	$C_B$	$\beta_U$	$\widehat{\beta}_u$	$\Delta_{MH}$	MH	SIB
1500	1500	$c$	0	.2	.026	.032	.27	.46	.58
1000	3000	0	0	.2	.032	.042	.27	.64	.70
3000	3000	0	0	.2	.032	.035	.27	.91	.95
1500	1500	$c$	.5	.2	.029	.035	.27	.51	.60
1000	3000	0	.5	.2	.034	.044	.27	.65	.72
3000	3000	0	.5	.2	.034	.038	.27	.91	.94
1500	1500	0	0	.3	.048	.052	.40	.84	.90
1000	3000	$c$	0	.3	.042	.053	.40	.87	.91
3000	3000	$c$	0	.3	.042	.045	.40	.97	1.00
1500	1500	0	.5	.3	.050	.047	.40	.99	.99
1000	3000	$c$	.5	.3	.042	.054	.40	.80	.84
3000	3000	$c$	.5	.3	.042	.064	.40	.91	.92

Table 10: Bias,  $a_7 = 0.4$ , ACT,  $n_b = 3$ ,  $\alpha = 0.05$ .

$J_F$	$J_R$	$c$	$d_T$	$C_B$	$\beta_U$	$\widehat{\beta}_u$	SIB
1500	1500	0	0	.2	.063	.069	.70
1000	3000	$c$	0	.2	.053	.067	.68
3000	3000	$c$	0	.2	.053	.053	.80
1500	1500	$c$	.5	.2	.055	.071	.60
1000	3000	0	.5	.2	.065	.083	.72
3000	3000	0	.5	.2	.065	.074	.96
1500	1500	0	0	.3	.093	.095	.91
1000	3000	0	0	.3	.093	.11	.89
3000	3000	$c$	0	.3	.080	.081	.99
1500	1500	0	.5	.3	.097	.12	.97
1000	3000	$c$	.5	.3	.084	.11	.89
3000	3000	$c$	.5	.3	.083	.09	1.00

Table 11: Bias,  $a_\eta = 0.8$ , ASVAB,  $n_b = 1$ ,  $\alpha = 0.05$ .

$J_F$	$J_R$	$c$	$d_T$	$C_\beta$	$\beta_U$	$\widehat{\beta}_U$	$\Delta_{MH}$	$MH$	$SIB$
1500	1500	c	0	.2	.026	.029	.27	.42	.50
1000	3000	0	0	.2	.034	.039	.27	.63	.79
3000	3000	0	0	.2	.034	.034	.27	.90	.95
1500	1500	c	.5	.2	.027	.035	.27	.63	.66
1000	3000	0	.5	.2	.034	.038	.27	.63	.70
3000	3000	0	.5	.2	.034	.036	.27	.89	.91
1500	1500	0	0	.3	.051	.052	.40	.85	.92
1000	3000	c	0	.3	.042	.044	.40	.77	.84
3000	3000	c	0	.3	.042	.046	.40	.99	.99
1500	1500	0	.5	.3	.051	.057	.40	.91	.93
1000	3000	c	.5	.3	.038	.048	.40	.77	.82
3000	3000	c	.5	.3	.039	.045	.40	.94	.97

Table 12: Bias,  $a_\eta = 0.4$ , ASVAB,  $n_b = 3$ ,  $\alpha = 0.05$ .

$J_F$	$J_R$	$c$	$d_T$	$C_\beta$	$\beta_U$	$\widehat{\beta}_U$	$SIB$
1500	1500	0	0	.2	.065	.067	.70
1000	3000	c	0	.2	.052	.056	.53
3000	3000	c	0	.2	.052	.053	.85
1500	1500	c	.5	.2	.052	.068	.63
1000	3000	0	.5	.2	.064	.083	.73
3000	3000	0	.5	.2	.064	.072	.92
1500	1500	0	0	.3	.098	.10	.94
1000	3000	0	0	.3	.097	.10	.97
3000	3000	c	0	.3	.079	.079	.98
1500	1500	0	.5	.3	.097	.011	.98
1000	3000	c	.5	.3	.076	.098	.87
3000	3000	c	.5	.3	.078	.090	.99

Distribution List

Dr. Terry Ackerman  
Educational Psychology  
210 Education Bldg.  
University of Illinois  
Champaign, IL 61801

Dr. James Algina  
1403 Norman Hall  
University of Florida  
Gainesville, FL 32605

Dr. Erling B. Andersen  
Department of Statistics  
Sønderstrøde 6  
1455 Copenhagen  
DENMARK

Dr. Ronald Armstrong  
Rutgers University  
Graduate School of Management  
Newark, NJ 07102

Dr. Eva L. Baker  
UCLA Center for the Study  
of Evaluation  
145 Moore Hall  
University of California  
Los Angeles, CA 90024

Dr. Laura L. Barnes  
College of Education  
University of Toledo  
2801 W. Bancroft Street  
Toledo, OH 43606

Dr. William M. Barr  
University of Minnesota  
Dept. of Educ. Psychology  
330 Burton Hall  
173 Pillsbury Dr., S.E.  
Minneapolis, MN 55455

Dr. Isaac Bejar  
Law School Admissions  
Services  
P.O. Box 40  
Newtown, PA 18940-0040

Dr. Ira Bernstein  
Department of Psychology  
University of Texas  
P.O. Box 19528  
Arlington, TX 76019-0528

Dr. Menucha Birenbaum  
School of Education  
Tel Aviv University  
Ramat Aviv 69978  
ISRAEL

Dr. Arthur S. Blahos  
Code N712  
Naval Training Systems Center  
Orlando, FL 32813-7100

Dr. Bruce Blomox  
Defense Manpower Data Center  
99 Pacific St.  
Suite 155A  
Monterey, CA 93943-3231

Cdt. Arnold Bohrer  
Seattle Psychologisch Onderzoek  
F. Struikings-Ea Selectiecentrum  
Kwartier Koningin Astrid  
Bruijstraat  
1120 Brussels, BELGIUM

Dr. Robert Brasuz  
Code 281  
Naval Training Systems Center  
Orlando, FL 32826-3224

Dr. Robert Brann  
American College Testing  
Programs  
P. O. Box 148  
Iowa City, IA 52243

Dr. Gregory Candell  
CTB/McGraw-Hill  
2500 Garden Road  
Monterey, CA 93940

Dr. John B. Carroll  
409 Elliott Rd., North  
Chapel Hill, NC 27514

Dr. John M. Carroll  
IBM Watson Research Center  
User Interface Institute  
P.O. Box 704  
Yorktown Heights, NY 10598

Dr. Robert M. Carroll  
Chief of Naval Operations  
OP-01B2  
Washington, DC 20350

Dr. Raymond E. Christal  
UES LAMP Science Advisor  
AFHRL/MOEL  
Brooks AFB, TX 78235

Mr. Hua Hua Chung  
University of Illinois  
Department of Statistics  
101 Illini Hall  
725 South Wright St.  
Champaign, IL 61820

Dr. Norman Cliff  
Department of Psychology  
Univ. of So. California  
Los Angeles, CA 90089-1061

Director, Manpower Program  
Center for Naval Analyses  
4401 Ford Avenue  
P.O. Box 16268  
Alexandria, VA 22302-0268

Director,  
Manpower Support and  
Readiness Program  
Center for Naval Analyses  
2000 North Bascom Street  
Alexandria, VA 22311

Dr. Stanley Colyer  
Office of Naval Technology  
Code 222  
800 N. Quincy Street  
Arlington, VA 22217-5000

Dr. Hans F. Crombag  
Faculty of Law  
University of Limburg  
P.O. Box 616  
Maastricht  
The NETHERLANDS 6200 MD

Ms. Carolyn R. Crone  
Johns Hopkins University  
Department of Psychology  
Charles & 34th Street  
Baltimore, MD 21218

Dr. Timothy Devey  
American College Testing Program  
P.O. Box 148  
Iowa City, IA 52243

Dr. C. M. Dayton  
Department of Measurement  
Statistics & Evaluation  
College of Education  
University of Maryland  
College Park, MD 20742

Dr. Ralph J. DeAyala  
Measurement, Statistics,  
and Evaluation  
Benjamin Bldg., Rm. 4112  
University of Maryland  
College Park, MD 20742

Dr. Lou DiBello  
CERL  
University of Illinois  
103 South Mathews Avenue  
Urbana, IL 61801

Dr. Dattaprasad Divgi  
Center for Naval Analyses  
4401 Ford Avenue  
P.O. Box 16268  
Alexandria, VA 22302-0268

Mr. Hai-Ki Dong  
Bell Communications Research  
Room PYA-1K207  
P.O. Box 1320  
Piscataway, NJ 08855-1320

Dr. Fritz Dragow  
University of Illinois  
Department of Psychology  
603 E. Daniel St.  
Champaign, IL 61820

Defense Technical  
Information Center  
Cameron Station, Bldg 5  
Alexandria, VA 22314  
(2 Copies)

Dr. Stephen Dunbar  
244B Lindquist Center  
for Measurement  
University of Iowa  
Iowa City, IA 52242

Dr. James A. Earles  
Air Force Human Resources Lab  
Brooks AFB, TX 78235

Dr. Susan Embretson  
University of Kansas  
Psychology Department  
426 Frazer  
Lawrence, KS 66045

Dr. George Englehard, Jr.  
Division of Educational Studies  
Emory University  
210 Fieburne Bldg.  
Atlanta, GA 30322

ERIC Facility-Acquisitions  
2440 Research Blvd, Suite 550  
Rockville, MD 20850-3238

Dr. Benjamin A. Fairbank  
Operational Technologies Corp.  
5425 Callaghan, Suite 225  
San Antonio, TX 78228

Dr. Marshall J. Farr, Consultant  
Cognitive & Instructional Sciences  
2520 North Vernon Street  
Arlington, VA 22207

Dr. P.A. Federico  
Code 51  
NPRDC  
San Diego, CA 92152-6800

Dr. Leonard Feldt  
Lindquist Center  
for Measurement  
University of Iowa  
Iowa City, IA 52242

Dr. Richard L. Ferguson  
American College Testing  
P.O. Box 168  
Iowa City, IA 52243

Dr. Gerhard Fischer  
Liebigasse 5/3  
A 1016 Vienna  
AUSTRIA

Dr. Myron Fiachl  
U.S. Army Headquarters  
DAPE-MRR  
The Pentagon  
Washington, DC 20310-0300

Prof. Donald Fitzgerald  
University of New England  
Department of Psychology  
Armidale, New South Wales 2351  
AUSTRIA

Mr. Paul Foley  
Navy Personnel R&D Center  
San Diego, CA 92152-6800

Dr. Alfred R. Freghy  
AFOSR/NL, Bldg. 410  
Bolling AFB, DC 20332-6448

Dr. Robert D. Gibbons  
Illinois State Psychiatric Inst.  
Rm 529W  
1601 W. Taylor Street  
Chicago, IL 60612

Dr. Janice Gifford  
University of Massachusetts  
School of Education  
Amherst, MA 01003

Dr. Drew Glomer  
Educational Testing Service  
Princeton, NJ 08541

Dr. Robert Glaser  
Learning Research  
& Development Center  
University of Pittsburgh  
3939 O'Hara Street  
Pittsburgh, PA 15260

Dr. Sherris Gott  
AFHRL/MOMJ  
Brooks AFB, TX 78235-5401

Dr. Bert Green  
Johns Hopkins University  
Department of Psychology  
Charles & 34th Street  
Baltimore, MD 21218

Michael Habon  
DORNIER GMBH  
P.O. Box 1420  
D-7990 Friedrichshafen 1  
WEST GERMANY

Prof. Edward Haertel  
School of Education  
Stanford University  
Stanford, CA 94305

Dr. Ronald K. Hambleton  
University of Massachusetts  
Laboratory of Psychometrics  
and Evaluative Research  
Hills South, Room 152  
Amherst, MA 01003

Dr. Delwyn Harnisch  
University of Illinois  
51 Gerry Drive  
Champaign, IL 61820

Dr. Grant Henning  
Senior Research Scientist  
Division of Measurement  
Research and Services  
Educational Testing Service  
Princeton, NJ 08541

Ms. Rebecca Hettler  
Navy Personnel R&D Center  
Code 63  
San Diego, CA 92152-6800

Dr. Thomas M. Hinech  
ACT  
P. O. Box 168  
Iowa City, IA 52243

Dr. Paul W. Holland  
Educational Testing Service, 21-T  
Rosedale Road  
Princeton, NJ 08541

Dr. Paul Horst  
677 G Street, #184  
Chula Vista, CA 92010

Ms. Julia S. Hough  
Cambridge University Press  
40 West 20th Street  
New York, NY 10011

Dr. William Howell  
Chief Scientist  
AFHRL/CA  
Brooks AFB, TX 78235-5401

Dr. Lloyd Humphreys  
University of Illinois  
Department of Psychology  
603 East Daniel Street  
Champaign, IL 61820

Dr. Steven Hunka  
3-104 Educ. N.  
University of Alberta  
Edmonton, Alberta  
CANADA T6G 2G5

Dr. Huynh Huynh  
College of Education  
Univ. of South Carolina  
Columbia, SC 29208

Dr. Robert Jaanarone  
Elec. and Computer Eng. Dept.  
University of South Carolina  
Columbia, SC 29208

Dr. Kumar Jog-dev  
University of Illinois  
Department of Statistics  
101 Illini Hall  
725 South Wright Street  
Champaign, IL 61820

Dr. Douglas H. Jones  
1280 Woodfern Court  
Toms River, NJ 08753

Dr. Brian Junker  
Carnegie-Mellon University  
Department of Statistics  
Schenley Park  
Pittsburgh, PA 15213

Dr. Michael Kaplan  
Office of Basic Research  
U.S. Army Research Institute  
5001 Eisenhower Avenue  
Alexandria, VA 22333-5400

Dr. Milton S. Katz  
European Science Coordination  
Office  
U.S. Army Research Institute  
Box 65  
FPO New York 09510-1500

Prof. John A. Kasts  
Department of Psychology  
University of Newcastle  
N.S.W. 2308  
AUSTRALIA

Dr. Jun-keun Kim  
Department of Psychology  
Middle Tennessee State  
University  
P.O. Box 522  
Murfreesboro, TN 37132

Mr. Soon-Hoon Kim  
Computer-based Education  
Research Laboratory  
University of Illinois  
Urbana, IL 61801

Dr. G. Gage Kingsbury  
Portland Public Schools  
Research and Evaluation Department  
501 North Dixon Street  
P. O. Box 3107  
Portland, OR 97209-3107

Dr. William Koch  
Box 7246, Meas. and Eval. Ctr.  
University of Texas-Austin  
Austin, TX 78703

Dr. Richard J. Koubek  
Department of Biomedical  
& Human Factors  
139 Engineering & Math Bldg.  
Wright State University  
Dayton, OH 45435

Dr. Leonard Kroeker  
Navy Personnel R&D Center  
Code 62  
San Diego, CA 92152-6800

Dr. Jerry Lehnus  
Defense Manpower Data Center  
Suite 400  
1600 Wilson Blvd  
Rosslyn, VA 22209

Dr. Thomas Leonard  
University of Wisconsin  
Department of Statistics  
1210 West Dayton Street  
Madison, WI 53705

Dr. Michael Levine  
Educational Psychology  
210 Education Bldg.  
University of Illinois  
Champaign, IL 61801

Dr. Charles Lewis  
Educational Testing Service  
Princeton, NJ 08541-0001

Mr. Rodney Lin  
University of Illinois  
Department of Psychology  
603 E. Daniel St.  
Champaign, IL 61820

Dr. Robert L. Lina  
Campus Box 249  
University of Colorado  
Boulder, CO 80309-0249

Dr. Robert Lockman  
Center for Naval Analysis  
4401 Ford Avenue  
P.O. Box 16368  
Alexandria, VA 22302-0268

Dr. Frederic M. Lord  
Educational Testing Service  
Princeton, NJ 08541

Dr. Richard Luecht  
ACT  
P. O. Box 168  
Iowa City, IA 52243

Dr. George B. Macready  
Department of Measurement  
Statistics & Evaluation  
College of Education  
University of Maryland  
College Park, MD 20742

Dr. Gary Marco  
Step 31-E  
Educational Testing Service  
Princeton, NJ 08541

Dr. Clasen J. Martin  
Office of Chief of Naval  
Operations (OP 13 F)  
Navy Annex Room 2832  
Washington, DC 20350

Dr. James R. McBride  
HumRRO  
6430 Elmburst Drive  
San Diego, CA 92120

Dr. Clarence C. McCormick  
HQ, USMEPCOM/MEPCT  
2500 Green Bay Road  
North Chicago, IL 60064

Mr. Christopher McCusker  
University of Illinois  
Department of Psychology  
603 E. Daniel St.  
Champaign, IL 61820

Dr. Robert McKinley  
Educational Testing Service  
Princeton, NJ 08541

Mr. Alan Mead  
c/o Dr. Michael Levine  
Educational Psychology  
210 Education Bldg.  
University of Illinois  
Champaign, IL 61801

Dr. Timothy Miller  
ACT  
P. O. Box 168  
Iowa City, IA 52243

Dr. Robert Mislevy  
Educational Testing Service  
Princeton, NJ 08541

Dr. William Montague  
NPRDC Code 13  
San Diego, CA 92152-6800

Ms. Kathleen Moreno  
Navy Personnel R&D Center  
Code 62  
San Diego, CA 92152-6800

Headquarters Marine Corps  
Code MPI-20  
Washington, DC 20380

Dr. Ratna Nandakumar  
Educational Studies  
Ward Hall, Room 213E  
University of Delaware  
Newark, DE 19716

Library, NPRDC  
Code P201L  
San Diego, CA 92152-6800

Literates  
Naval Center for Applied Research  
in Artificial Intelligence  
Naval Research Laboratory  
Code 5510  
Washington, DC 20375-5000

Dr. Harold F. O'Neil, Jr.  
School of Education - WPH 801  
Department of Educational  
Psychology & Technology  
University of Southern California  
Los Angeles, CA 90089-0031

Dr. James B. Olsen  
WICAT Systems  
1875 South State Street  
Oram, UT 84058

Office of Naval Research,  
Code 1142CS  
800 N. Quincy Street  
Arlington, VA 22217-5000  
(6 Copies)

Dr. Judith Orasanu  
Basic Research Office  
Army Research Institute  
5001 Eisenhower Avenue  
Alexandria, VA 22333

Dr. Jesse Orfanaky  
Institute for Defense Analyses  
1801 N. Beauregard St.  
Alexandria, VA 22311

Dr. Peter J. Pashley  
Educational Testing Service  
Rosedale Road  
Princeton, NJ 08541

Wayne M. Patience  
American Council on Education  
GED Testing Service, Suite 20  
One Dupont Circle, NW  
Washington, DC 20036

Dr. James Paulson  
Department of Psychology  
Portland State University  
P.O. Box 751  
Portland, OR 97207

Dept. of Administrative Sciences  
Code 54  
Naval Postgraduate School  
Monterey, CA 93943-5026

Dr. Mark D. Reckase  
ACT  
P. O. Box 168  
Iowa City, IA 52243

Dr. Malcolm Ree  
AFHRL/MOA  
Brooks AFB, TX 78235

Mr. Steve Reiss  
N660 Elliott Hall  
University of Minnesota  
75 E. River Road  
Minneapolis, MN 55455-0344

Dr. Carl Ross  
CNET-PDCD  
Building 90  
Great Lakes NTC, IL 60088

Dr. J. Ryan  
Department of Education  
University of South Carolina  
Columbia, SC 29208

Dr. Fumiko Semejima  
Department of Psychology  
University of Tennessee  
310B Austin Peay Bldg.  
Knoxville, TN 37916-0900

Mr. Drew Sands  
NPRDC Code 62  
San Diego, CA 92152-6800

Lowell Sobocor  
Psychological & Quantitative  
Foundations  
College of Education  
University of Iowa  
Iowa City, IA 52242

Dr. Mary Schratz  
4100 Parlaide  
Carlsbad, CA 92008

Dr. Dan Segall  
Navy Personnel R&D Center  
San Diego, CA 92152

Dr. Robin Shealy  
University of Illinois  
Department of Statistics  
101 Illini Hall  
725 South Wright St.  
Champaign, IL 61820

Dr. Kazuo Shigemasa  
7-9-24 Kugenuma-Kaigan  
Fujisawa 251  
JAPAN

Dr. Randall Shumaker  
Naval Research Laboratory  
Code 5510  
4555 Overlook Avenue, S.W.  
Washington, DC 20375-5000

Dr. Richard E. Soow  
School of Education  
Stanford University  
Stanford, CA 94305

Dr. Richard C. Sorenson  
Navy Personnel R&D Center  
San Diego, CA 92152-6800

Dr. Judy Spray  
ACT  
P.O. Box 168  
Iowa City, IA 52243

Dr. Martha Stocking  
Educational Testing Service  
Princeton, NJ 08541

Dr. Peter Stuloff  
Center for Naval Analysis  
4401 Ford Avenue  
P.O. Box 14268  
Alexandria, VA 22302-0268

Dr. William Stout  
University of Illinois  
Department of Statistics  
101 Illini Hall  
725 South Wright St.  
Champaign, IL 61820

Dr. Haribaran Swaminathan  
Laboratory of Psychometric and  
Evaluation Research  
School of Education  
University of Massachusetts  
Amherst, MA 01003

Mr. Brad Symeon  
Navy Personnel R&D Center  
Code-62  
San Diego, CA 92152-6800

BEST COPY AVAILABLE

Dr. John Tangney  
AFOSR/NL, Bldg. 410  
Bolling AFB, DC 20332-6448

Dr. Kikumi Tatsuoka  
Educational Testing Service  
Mail Stop 03-T  
Princeton, NJ 08541

Dr. Maurice Tatsuoka  
Educational Testing Service  
Mail Stop 03-T  
Princeton, NJ 08541

Dr. David Thlesen  
Department of Psychology  
University of Kansas  
Lawrence, KS 66044

Mr. Thomas J. Thomas  
Johns Hopkins University  
Department of Psychology  
Charles & 34th Street  
Baltimore, MD 21218

Mr. Gary Thomason  
University of Illinois  
Educational Psychology  
Champaign, IL 61820

Dr. Robert Tsutakawa  
University of Missouri  
Department of Statistics  
222 Math. Sciences Bldg.  
Columbia, MO 65211

Dr. Ledyard Tucker  
University of Illinois  
Department of Psychology  
603 E. Daniel Street  
Champaign, IL 61820

Dr. David Vale  
Assessment Systems Corp.  
2233 University Avenue  
Suite 440  
St. Paul, MN 55114

Dr. Frank L. Vicino  
Navy Personnel R&D Center  
San Diego, CA 92152-6800

Dr. Howard Wainer  
Educational Testing Service  
Princeton, NJ 08541

Dr. Michael T. Walker  
University of Wisconsin-Milwaukee  
Educational Psychology Department  
Box 413  
Milwaukee, WI 53201

Dr. Ming-Mei Wang  
Educational Testing Service  
Mail Stop 03-T  
Princeton, NJ 08541

Dr. Thomas A. Warm  
FAA A-400y AAC934D  
P.O. Box 25082  
Oklahoma City, OK 73125

Dr. Brian Waters  
HumRRO  
1100 S. Washington  
Alexandria, VA 22314

Dr. David J. Weiss  
N660 Elliott Hall  
University of Minnesota  
75 E. River Road  
Minneapolis, MN 55455-0344

Dr. Ronald A. Weitzman  
Box 146  
Carmel, CA 93921

Major John Welsh  
AFHRL/MOAN  
Brooks AFB, TX 78223

Dr. Douglas Wetzel  
Code 51  
Navy Personnel R&D Center  
San Diego, CA 92152-6800

Dr. Rand R. Wilcox  
University of Southern  
California  
Department of Psychology  
Los Angeles, CA 90089-1061

German Military Representative  
ATTN: Wolfgang Wildgrube  
Stratkräfteamt  
D-5300 Bonn 2  
4000 Brandywine Street, NW  
Washington, DC 20016

Dr. Bruce Williams  
Department of Educational  
Psychology  
University of Illinois  
Urbana, IL 61801

Dr. Hilda Wing  
Federal Aviation Administration  
800 Independence Ave. SW  
Washington, DC 20591

Mr. John H. Wolfe  
Navy Personnel R&D Center  
San Diego, CA 92152-6800

Dr. George Wong  
Biostatistics Laboratory  
Memorial Sloan-Kettering  
Cancer Center  
1275 York Avenue  
New York, NY 10021

Dr. Wallace Wulfert, III  
Navy Personnel R&D Center  
Code 51  
San Diego, CA 92152-6800

Dr. Kentaro Yamamoto  
02-T  
Educational Testing Service  
Rosedale Road  
Princeton, NJ 08541

Dr. Wendy Yen  
CTB/McGraw Hill  
Del Monte Research Park  
Moorpark, CA 93940

Dr. Joseph L. Young  
National Science Foundation  
Room 320  
1800 G Street, N.W.  
Washington, DC 20550

Mr. Anthony R. Zarn  
National Council of State  
Boards of Nursing, Inc.  
625 North Michigan Avenue  
Suite 1544  
Chicago, IL 60611