DOCUMENT RESUME

ED 347 170 TM 018 528

AUTHOR Sykes, Robert C.; And Others
TITLE Assessing the Impact of Multidimensionality on the Classification Decisions of an IRT-Based Licensure Examination.
PUB DATE Apr 92
NOTE 27p.; Paper presented at the Annual Meeting of the National Council on Measurement in Education (San Francisco, CA, April 21-23, 1992).
PUB TYPE Reports - Evaluative/Feasibility (142) -- Speeches/Conference Papers (150)

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS *Classification; Comparative Testing; Computer Simulation; *Decision Making; Failure; Higher Education; *Item Response Theory; *Licensing Examinations (Professions); Pass Fail Grading; *Scores; Test Format
IDENTIFIERS *Multidimensionality (Tests); *Part Form Method; Rasch Model

ABSTRACT

A part-form methodology was used to study the effect of varying degrees of multidimensionality on the consistency of pass/fail classification decisions obtained from simulated unidimensional item response theory (IRT) based licensure examinations. A control on the degree of form multidimensionality permitted an assessment throughout the range of multidimensionality of any potential effect on Rasch item parameters and pass/fail classifications obtained from scores derived from them. Four full-length (300-item) forms of a licensure examination produced by CTB Macmillan/McGraw Hill were used to generate part-forms for four administrations in the summer of 1988, winter of 1989, summer of 1989, and winter of 1990, respectively. There were 2,000 examinees for each form. All four full-length forms had been demonstrated to be multidimensional, but could be made unidimensional by deleting no more than half the items with the largest absolute loadings on the second factor. Overall, failure concordance percentages did not differ between those pairs of part-forms that differed maximally in the degree of predicted multidimensionality and those pairs of part-forms where members were both predicted to be multidimensional. Results suggest that increased multidimensionality had no substantial effect on failure decision agreement. However, the failure concordance percentages for pairs of part-forms that were both most likely unidimensional were slightly higher than those for other pairs of part-forms. Four tables, 1 figure of study data and 12 references are included. (SLD)

ASSESSING THE IMPACT OF MULTIDIMENSIONALITY ON THE CLASSIFICATION DECISIONS OF AN IRT-BASED LICENSURE EXAMINATION

Robert C. Sykes
Kyoko Ito
Raylene Potter

CTB Macmillan/McGraw-Hill

This paper was presented in April, 1992 at the Annual Meeting of the National Council on Measurement in Education San Francisco, CA

Introduction

Although most of the currently applicable IRT models assume unidimensionality, "real life" test data are rarely unidimensional (Harrison, 1986; Humphreys, 1986; Linn, Levine, Hastings, & Wardrop, 1981; Traub, 1983). The inherently multidimensional nature of real data and the importance of the assumption of unidimensionality for measurement has resulted in the development of various methods of testing for unidimensionality (Hattie, 1985).

In addition to these methods of evaluating dimensionality, substantial attention has been devoted to evaluating the consequences of violating the unidimensionality assumption. In relation to the robustness of item response theory (IRT) models, the studies by Drasgow and Parsons (1983) and Harrison (1986) have shown that item and trait parameters implied by a second-order, general factor are recovered effectively when the intercorrelations between common factors are at least moderate.

Several studies have shown that multidimensionality can manifest itself as differential item functioning (dif). For example, Ackerman (1988) has demonstrated that the application of a unidimensional IRT model to two-dimensional data can result in dif if the multidimensional ability distributions are unequal between groups. Using simulated data, Oshima and Miller (1991) have shown that, irrespective of whether groups differ on the trait of concern, a small percentage of items that are

multidimensional and biased can be correctly differentiated from a set of multidimensional but unbiased items. After discussing the problems associated with some IRT item bias detection procedures (i.e., the separate calibration approach and the combined reference and focal group approach), Wang (1988) suggests the IRT-likelihood ratio method proposed by Thissen, Steinberg, and Wainer (1988) as a best approach to correctly detecting dif as an indication of multidimensionality.

Other studies have investigated the effect of multidimensionality on IRT-based ability estimates. Using a dataset with the ten easiest items measuring one trait and the ten hardest items measuring another trait, Reckase, Carlson, Ackerman, and Spray (1986) have shown that the confounding of item difficulty and dimensionality can cause the unidimensional ability scale to have a different meaning at different points on the scale. Reckase (1979) has found that stable IRT-based ability estimates from multidimensional data require that the first factor accounts for at least 20 percent of the test variance.

Thus, research on the effects of multidimensionality has been directed toward issues such as dif and parameter estimates and not toward the effects on classification decisions generated from examination forms. The consequences of multidimensionality on pass/fail classifications are of particular significance to licensure and certification examinations that often generate a single score that is used to determine whether or not an examinee

can enter the profession. As Reckase (1979) has pointed out, satisfying the unidimensionality assumption is more of a problem with licensure and certification examinations and achievement tests than with psychological tests. Factor-pure psychological tests can be more easily constructed using traditional item analysis and factor analysis procedures than licensure/certification examinations and achievement tests. Licensure examinations are constructed according to the test plan specifications and usually do not rely on factor-analytic procedures.

It must be noted that while the legal defensibility of licensure examinations has rested primarily on the items in each examination representing a test plan based upon a job analysis, the validity of these examinations as well as others that are administered as multiple forms over time requires that scores should also be comparable over testing occasions. An examinee's score should not depend upon which test plan representative form was taken.

In order for scores to be comparable, the significant dimensionality of examinations must be known and subsequently, like specific content representation, fixed over forms. When the dimensionality of examinees' item performance coincides with the specified content categories, additional content constraints are not required. The fact that item performance, while multidimensional, is coincident with content categories does not, however, necessarily guarantee that scores will be comparable if

based upon an IRT model assuming unidimensionality. If the dimensionality or factor structure of examinations is not known or does not coincide with the specified content structure, content cannot be presumed to be adequately controlled.

Under these circumstances, performance on examinations may be determined by different dimensions on different occasions. This may have consequences for the adequacy of item parameters that have been generated from an IRT model assuming unidimensionality or for examinee scores produced from them. Consequently, for licensure/certification examinations that are administered in multiple forms, it is important to investigate the consistency of pass/fail classifications over different forms of an examination that may differ in dimensionality.

In this study, a part-form methodology was used to study the effect of varying degrees of multidimensionality on the consistency of pass/fail classification decisions obtained from simulated, unidimensional IRT-based, licensure examinations. A control on the degree of form multidimensionality permitted an assessment throughout the range of multidimensionality of any potential effect on Rasch item parameters and pass/fail classifications obtained from scores derived from them.

Method

Full-Length Forms

Four full-length (300-item) forms of a licensure examination produced by CTB were used to generate part-forms.[1] The forms had been constructed according to the content category quotas specified by a test plan that was based on a recent job analysis. The test plan specifies minimum and maximum quotas for each of a number of categories in the two content domains. Dates of administration for the four forms were Summer 1988, Winter 1989, Summer 1989, and Winter 1990. The forms are denoted in this paper as 288, 189, 289, and 190, respectively. The items in the forms had been calibrated using the Rasch model. Full-length forms of the licensure examination had been demonstrated to be unspeeded on a number of past occasions.

Construction of Part-Forms

Part-forms were created utilizing items from each of the full-length forms. A methodology of creating part-forms was adopted because it allowed comparison of classification decisions across a large number of forms that could vary in dimensionality.

Part-forms varied in form length and the degree of predicted multidimensionality. The part-forms of primary interest were "quarter forms" containing approximately 75 items each that were predicted to be either purely unidimensional on the ability of

---

[1] Because of negative point biserials, one item was deleted from each of two forms (288 and 189). Those forms had 299 scored items, instead of 300 items.

concern (i.e., 0% multidimensional) or almost certainly multidimensional (i.e., 100% multidimensional). Two other form sizes were also examined: "half-forms" that contained approximately 150 items and "third-forms" that consisted of approximately 100 items.

Predicted multidimensionality of a part-form (referred to as "percent multidimensionality" (% md) in this paper) was determined based on the second-factor loadings obtained from the dimensionality analyses of the full-length forms performed using the Stout procedure (Stout, 1987, 1990). The Stout procedure tests local independence by comparing a unidimensional variance estimate to a theoretical or usual variance estimate on a part-test that has the greatest chance of being multidimensional. "Percent multidimensionality" refers to the percentage of a part-form constituted by items selected from the 50 percent of the items having the largest, absolute-valued, second-factor loadings in the Stout analysis of the full-length form from which the items were drawn. The operational definition of part-form multidimensionality was derived based on the fact that every form studied was originally multidimensional but could be made unidimensional, while retaining test plan representativeness, by deleting no more than the one half of its items that loaded most heavily on the second factor. Quarter forms that are 0% md then are predicted most likely to be unidimensional and forms that are 100% md predicted to be most likely multidimensional.

Table 1 presents the results of the Stout analyses of the

full-length forms from which various part-forms were constructed. The two 1989 forms (189 and 289) and the 1990 form (190) proved to be clearly multidimensional ($p < .01$), based upon the Stout assessment. The 288 form was marginally unidimensional ($p = .09$).

Eigenvalues from the factor analyses accompanying the Stout procedure suggested that the multidimensionality could be attributed to the existence of only one other salient factor. Evaluation of these eigenvalue differences as well as those available from other analyses revealed that a difference between the second and third eigenvalues greater than .600 was always associated with a multidimensional Stout statistic. On the other hand, a difference between the second and third eigenvalues that was less than .500 was always associated with a unidimensional Stout statistic. Differences between .500 and .600 could be associated with either a multidimensional or unidimensional statistic.

Based upon a prepotency of the second factor in inducing multidimensionality, items with large absolute-valued second factor loadings in each of the forms were assigned to part-forms. The following example with a set of two 100% md and two 0% md quarter forms illustrates how these items in the 288 form were allocated to the part-forms. First, all 300 items in the full-length form were sorted in descending order of absolute-valued second-factor loadings. Second, the 50 percent of the items having the largest second-factor loadings in absolute magnitude

were allocated to the two 100% md forms in such a way that all of the items in those quarter forms consisted of these high-loading items. None of the items in the two 0% md quarter forms came from the high-loading group.

The 50 percent of the items with the largest (absolute) second factor loadings in a full-length form were allocated to a set of part-forms in a systematic fashion. In the set described above (containing two 100% and two 0% md quarter forms), one of every two items in the high-loading group was assigned to the first 100% md quarter form, and the other item in the pair to the second 100% md quarter form. To avoid always assigning the first, higher-loading item of every two items to the first 100% md form, the two 100% quarter forms alternately received the high-loading items. For example, in the first round of allocation, the first 100% md form was allocated a high-loading item first, and the second 100% md form received the next highest loading item. In the second round, the second 100% md form received the third highest loading item, and the first 100% md form received the fourth ranked item, and so forth.

As with the construction of full-length forms, special care was taken to ensure that the constructed part-forms fulfilled the test plan percentage quotas. The unspeeded nature of the full-length forms minimized the possibility of a confounding speededness dimension impacting performance on the part-forms.

Analyses

The effect of multidimensionality on pass/fail decisions was assessed by comparing failure concordance rates across pairings of 0%-0%, 0%-100%, 0%-100%, and 100%-100% md quarter forms. (Hereafter the md term will be dropped from the pair name; i.e., 100%-100%). A failure concordance rate for a pair of part-forms was defined as misclassification of failures, assuming that the latter member of the part-form pair generated "true" classification decisions. To illustrate, assume a 100%-0% pair of quarter forms in which Form 1 is 100% md and Form 2 is 0% md. Two concordance rates (i.e., percentages) on failure decisions were calculated for this pair in the following manner: The failure concordance rate for Form 1 (i.e., denoted as "100%-0%" in the tables) was computed by dividing N_4 by N_{3+4}. The corresponding rate for Form 2 (0%-100%) was obtained by dividing N_4 by N_{2+4}.

		Form 2 (0% md) Pass	Form 2 (0% md) Fail	
Form 1 (100% md)	Pass	N_1	N_2	N_{1+2}
	Fail	N_3	N_4	N_{3+4}
		N_{1+3}	N_{2+4}	N

Mean failure concordance rates were then computed by averaging over all pairs that had the same permutation of % md members. For example, a mean failure concordance rate was calculated for all 100%-0% pairs and similarly, a rate was computed for all 0%-100% pairs.

Failure concordance rates, as opposed to overall pass/fail concordance rates or success concordance rates, were studied for two reasons. First, the examination has a high passing rate. The passing rate for the examination typically is 85 percent or higher for a population of first-time, U.S.-educated candidates. With a high percentage of passing candidates, high overall or success concordance rates would be difficult to interpret.

To illustrate this, assume Form 1 has a passing rate of 85 percent. An overall or success concordance rate of 85 percent could be achieved by simply adopting the decision rule of passing everyone on Form 2.

Second, a low failure concordance rate means that one form tends to pass a relatively large number of candidates that fail the other form. From a public safety standpoint, licensing agencies might want to minimize the possibility that unqualified candidates pass a second examination, after failing the first, solely due to test unreliability.

Each part-form in a set was calibrated using the Rasch model and equated to the pool of licensure items. A raw score-to-theta table was then generated for the part-form on the basis of the

equated b-values, and the raw-score equivalent of the cut-off theta for the examination was determined. The cut-off score for a full-length form is a point on the theta scale. When the passing raw scores were set for all part-forms in the set, scores were obtained on each part-form for each candidate in a sample (2000 first-time, U.S.-educated candidates for each of the four forms), and the part-form passing raw scores were applied to the scores to determine the candidates' pass/fail status on the part-forms. Average failure concordance rates were then computed and used to determine the effect of multidimensionality on pass/fail classifications.

Additionally, the design allowed the comparison of Rasch b-values obtained from test plan representative part-forms differing in multidimensionality and length.

Results/Discussion

Descriptive statistics for the four sets of exam quarter forms are provided in Table 2. Only three test plan and difficulty representative quarter forms, one 100% md and two 0% md forms, could be constructed from the 189 form. This was due to the necessary deletion of several items that were very easy, and hence could not be subjected to the item tetrachoric factor analysis used in the Stout procedure and the test plan category representation of the 189 examination. One of the content categories had near the minimum number of items in the complete 189 form.

Also, because of the constraints arising from the test plan representation of the 288 form, one quarter form from that examination had one more item in one domain category than the test plan quota called for. That quarter form had 18.9% of its items from that domain category, .9% over the maximum allowable.

The quarter forms in Table 2 are very similar in difficulty, with average p-values ranging between .67 and .72. They are also comparable in terms of the dispersion of item difficulty within each form. The standard deviation of p-values ranges between .13 and .17.

The effect of systematically allocating items to quarter forms can be seen in the mean and standard deviation of absolute second factor loadings. The 100% md quarter forms have very similar mean second factor loadings - all .16 with the exception of the first 189 form (form #1) having a mean of .17 - and

consequently they all consistently differ substantially from the 0% md quarter forms, with their means ranging between .04 and .05. The standard deviation of the quarter form absolute second factor loadings is .07 for all 100% md forms and between .03 and .04 for the 0% md forms.

Each of the quarter forms from two of the four forms, 289 and 190, was evaluated by the Stout procedure in order to verify that the actual dimensionality of the forms agreed with the predicted. Stout T statistics and associated p-values for the eight quarter forms are given in Table 3. All 0% md forms were in fact, unidimensional (all p's $\geq$ .18) while three out of four 100% md quarter forms - the two 289 forms and first 190 form - were clearly multidimensional (all three p's $\leq$ .01). The second 190 100% md quarter form was marginally multidimensional (T = 1.50, p = .07).

The failure concordance rates (i.e., percentages) obtained by comparing failure decisions within the 21 pairs of quarter forms (2 x 21 = 42 percentages) are presented in Table 4. Concordance rates are listed by the four possible combinations or conditions (100%-100%, 100%-0%, 0%-100%, and 0%-0%) and are arranged such that the % md of the latter form (or pair member) varies within the % md of the former pair member. The latter pair member in each condition is considered to produce true failure decisions. This ordering facilitates a comparison of rates within pairs of conditions where the % md of one of the

members of the pair is held constant (e.g., 100%-100% vs 100%-0% and 0%-100% vs 0%-0%).

Within each condition-by-form cell (e.g., 100%-100% within the 288 form), quarter forms are compared in the left to right order that they are specified in Table 2. For the 100%-100% condition within the 288 form this means that the 43.80 rate is generated from comparing quarter form #1 versus quarter form #2 ("true") in Table 2, and the 60.83 rate results from the comparison of quarter form #2 versus form #1 ("true").[2]

The individual failure concordance rates in Table 4 demonstrate substantial variability, ranging between 42.18 and 61.52. The overall mean of the 42 rates was 52.63.

Failure concordance rates are averaged for each form by condition cell and marginal means provided for each condition (over forms) and for the combined 100%-0% and 0%-100% conditions. Both cell and marginal means appear in bold in Table 4 to the right of the concordance rates that are averaged.

A comparison of the cell means for each of the four forms as well as the marginal means for the 0%-0% versus 0%-100% and 100%-0% versus 100%-100% conditions indicates no signs of a deleterious effect of increased multidimensionality of one of the pair members on failure concordance rates. Between the 0%-0%

---

[2] Similarly, the 45.09 and 45.94 rates for the 100%-0% condition within the 288 form represent the comparison of quarter form #1 versus quarter form #3 and quarter form #1 versus quarter form #4 from Table 2. The 55.49 and 53.12 rates, separated by one blank line from the previous pair of rates, represent the comparison of quarter form #2 versus quarter form #3, then the comparison of quarter form #2 versus quarter form #4. Thus, a pair of rates not separated by a blank line was obtained for the same first pair member and two different second pair members. As a final example, the 56.27 and 49.87 rates for the 0%-100% condition within the 288 form would represent the failure concordance rates for quarter form #3 versus quarter form #1, then quarter form #3 versus quarter form #2.

and 0%-100% conditions, the average failure concordance rates increase for two forms (i.e., 52.92 to 54.28 for 288 and 46.33 to 55.40 for 190) and decrease for the other two forms. The marginal 0%-0% mean is almost identical to the marginal 0%-100% mean (54.10 vs 54.01, respectively). Similarly, two out of the three forms actually exhibit an increase in failure concordance rates between the 100%-0% and 100%-100% conditions (i.e., 288 and 190) and the marginal mean for the 100%-100% condition is actually slightly larger than that for the 100%-0% comparisons (52.55 versus 50.45, respectively).

There are no a priori reasons to expect means for the 0%-100% condition to differ significantly from those of the 100%-0% condition. The marginal means for the two conditions were 54.01 and 50.45 respectively. The difference of 3.56 between the two marginal means seems somewhat large, however, given the relatively large number of rates each marginal mean is based upon (14). The difference may be attributed to a low 44.79 mean failure concordance rate for the 100%-0% condition within the 190 form. If marginal 0%-100% and 100%-0% means are computed, excluding the 190 form, the difference between the conditions is reduced to 0.74 (53.45 for 0%-100% minus 52.71 for 100%-0%). The low 100%-0%, form 190 cell mean cannot be attributed to the marginal multidimensionality of the second 100% 190 quarter form (#2 in Table 3). The mean concordance rate for the 100%-0% comparisons for that quarter form (43.06 and 46.11) is 44.59 versus 44.99 for the average concordance rate for the two

comparisons involving the other 100% quarter form from the 190 form (47.79 and 42.18).

Collapsing over all 100%-0% and 0%-100% cell means, the marginal mean of 52.23 is very similar to the marginal mean for the 100%-100% condition (52.55). Hence there appears to be no difference between failure concordance rates for pairs of forms maximally differing in % md and pairs of forms that are both multidimensional. The 0%-0% marginal mean of 54.10 is slightly larger than the weighted mean of 52.29 over the combined 0%-100%, 100%-0%, and 100%-100% conditions. Three of the four 0%-0% cell means are also greater than the weighted mean over the 100%-100%, 100%-0%, and 0%-100% conditions within each of the three forms (52.92 versus 52.14 for 288, 56.52 versus 55.40 for 189, and 60.64 versus 52.40 for 289). The 0%-0% mean of 46.33 for the 190 form was, however, substantially below the 51.08 weighted mean for the other three conditions.

The analysis of differences in failure concordance rates across quarter forms that ranged between 0% and 100% in predicted multidimensionality, as defined by the operational definition of multidimensionality, suggested at most a slight decrease in failure concordance rates when pairs of forms that were both most likely unidimensional (i.e., 0% md) were compared to those that contained at least one 100% md form. Quarter forms that are 0% and 100% md, however, represent the extremes of predicted multidimensionality that might be expected to occur in forms of this size.

In order to get some sense of whether failure concordance rates differed across pairs of part-forms that differed less extremely than the 0% and 100% quarter forms in their % md, one of the four forms was used to construct test plan and difficulty representative third forms of 25%, 50%, and 75% md. After pairing up the three possible pairings of third forms - 25%-50%, 25%-75% and 50%-75% - the mean failure concordance rate for the 25%-75% pairing was approximately three percent lower than the mean failure concordance rate for the two pairings, 25%-50% and 50%-75%, having less extreme differences in % md (55.09 versus 58.01, respectively).

Effects on Item Parameters

Two types of investigations evaluating effects on Rasch b-values were performed. Effects of reducing test length on b-values were initially assessed holding constant % md. Second, the effects of varying part-form multidimensionality were evaluated holding size of part-form constant.

A pair of 0% and 100% md test plan and difficulty representative half forms was created from the 288 form. Items were calibrated in each half-form using the Rasch model and a second form sample, and b-values from the 100% md half form paired with Rasch b-values obtained from the 100% md quarter form. In a similar manner, pairs of b-values from the 288 items in the 0% md half form and 0% md quarter forms were paired.

Correlations of b-values within each of the two sets of paired b-values were above .990.

Additionally, a set of test plan and difficulty representative third forms wa', created from the 288 form. Each of the three 50% md third forms was calibrated (on a third 2000 candidate sample) and Rasch b-values for the third forms paired with the pool b-value estimates for the items. The pool b-values for the 288 form were generated when the form was administered in 1988. All three correlations were above .990.

The effects of variable degrees of predicted multidimensionality on Rasch b-values were evaluated at two different part-form lengths; quarter and half forms. An additional set of test plan and difficulty representative quarter forms was created from the 189 form and two new sets of test plan and difficulty representative half forms from the 190 form. Each new form was calibrated and b-values for the items paired up over part-forms that differed by 25% to 75% in predicted multidimensionality for the quarter forms and 25% in predicted multidimensionality for the half forms. All eight correlations were above .980.

Conclusion

The present study investigated the effects of multidimensionality on the consistency of failure decisions obtained from pairs of part-forms. Part-forms were constructed from four 300-item forms of an IRT-based licensure examination with each part-form containing one quarter of the items in the full-length form from which the items were drawn. All four 300-item forms had originally been demonstrated to be multidimensional by the Stout procedure but could be made unidimensional by deleting no more than half of the items having the largest absolute loadings on the second factor. All quarter forms were verified to meet the examination test plan and have similar average difficulties. Quarter forms of primary interest were either purely unidimensional (0% md) or almost certainly multidimensional (100% md). Four combinations of percent multidimensionality were studied: 0%-0%, 0%-100%, 100%-0%, and 100%-100%. Concordance rates (percentages) on failure decisions were examined because of the high passing rates of the examination and the importance of minimizing inconsistent failure decisions from a public safety standpoint.

Twenty-one pairs of quarter forms produced 42 failure concordance percentages, ranging between 42.18% and 61.52%. When the percentages were collapsed over all 100%-0% md and 0%-100% md pairs, the mean of 52.23% was very similar to that for the 100%-100% condition (52.55%). The mean failure concordance rate for the 0%-0% md pairs (54.10%) was slightly higher than that for

the combined 100%-100% and 100%-0% pairs.

Overall, failure concordance percentages did not differ between those pairs of part-forms that differed maximally in the degree of predicted multidimensionality and those pairs of part-forms whose members were both predicted to be multidimensional. These results suggest no substantial effect of increased multidimensionality on failure decision agreement. However, the failure concordance percentages for pairs of part-forms that were both most likely unidimensional (0%-0%) were slightly higher than other pairs of part-forms.

A few comparisons among pairs of part forms of a predicted multidimensionality more likely to be constructed also suggested that pairs of part-forms that were closer in predicted multidimensionality might have higher failure concordance percentages than those pairs that more greatly differed in predicted multidimensionality. The presence of such an effect among additional forms of a predicted multidimensionality more likely to be constructed needs to be verified.

No substantial effects of reducing test length or multidimensionality on Rasch b-values were found.

References

Drasgow, F., & Parsons, C. K. (1983). Application of unidimensional item response theory models to multidimensional data. Applied Psychological Measurement, 7, 189-199.

Harrison, D. (1986). Robustness of IRT parameter estimation to violations of the unidimensionality assumption. Journal of Educational Statistics, 11, 91-115.

Hattie, J. (1985). Methodology review: Assessing unidimensionality of tests and items. Applied Psychological Measurement, 9, 139-164.

Humphreys, L. G. (1986). An analysis and evaluation of test and item bias in the prediction context. Journal of Applied Psychology, 71, 327-333.

Linn, R. L., Levine, M. V., Hastings, C. N., & Wardrop, J. L. (1981). Item bias in a test of reading comprehension. Applied Psychological Measurement, 5, 159-173.

Oshima, T. C. & Miller, M. D. (1991, April). Multidimensionality and item bias in item response theory. Paper presented at the Annual Meeting of the American Educational Research Association, Chicago.

Reckase, M. D. (1979). Unifactor latent trait models applied to multifactor tests: Results and implications. Journal of Educational Statistics, 4, 207-230.

Reckase, M. D., Carlson, J. E., Ackerman, T. A., & Spray, J. A. (1986, June). The interpretation of unidimensional IRT parameters when estimated from multidimensional data. Paper presented at the meeting of the Psychometric Society, Toronto.

Stout, W. (1987). A nonparametric approach for assessing latent trait unidimensionality. Psychometrika, 52, 589-617.

Thissen, D., Steinberg, L., & Wainer, H. (1988). Use of item response theory in the study of group differences in trace lines. In Wainer, H. & Braun, H. I. (Eds.), Test Validity. Hillsdale, NJ: Lawrence Erlbaum.

Traub, R. E. (1983). A priori considerations in choosing an item response model. In R. K. Hambleton (Ed.), Applications of item response theory. Vancouver: Educational Research Institute of British Columbia.

Wang, M. (1988, April). Measurement bias in the application of a unidimensional model to multidimensional item-response data. Paper presented at the Annual Meeting of the American Educational Research Association, New Orleans.

Table 1

First 10 Eigenvalues from the Linear Factor Analyses of the Examinations Assessed for Dimensionality: 288, 189, 289 and 190

Full Form

288		189		289		190	
Eigenvalue	Difference	Eigenvalue	Difference	Eigenvalue	Difference	Eigenvalue	Difference
15.111	10.119	15.568	10.540	17.203	12.281	16.358	11.680
4.992	0.537	5.028	1.137	4.922	1.180	4.678	.517
4.455	0.380	3.892	.629	3.742	.414	4.161	.426
4.075	0.180	3.263	.157	3.328	.155	3.736	.250
3.895	0.267	3.106	.152	3.173	.064	3.485	.112
3.628	0.152	2.954	.059	3.109	.198	3.373	.159
3.476	0.052	2.894	.021	2.911	.084	3.215	.148
3.424	0.050	2.873	.199	2.827	.042	3.066	.089
3.374	0.067	2.674	.045	2.785	.035	2.977	.054
3.307	0.088	2.629	.100	2.751	.061	2.923	.046
T = 1.33 n.s. (p = .09)		T = 3.90 sign. (p < .01)		T = 3.61 sign. (p < .01)		T = 2.73 sign. (p < .01)	

sign. = significant

Table 2

Descriptive Statistics for the Quarter Forms

	Exam															
	288				189				289				190			
	1	2	3	4	1	2	3		1	2	3	4	1	2	3	4
Size Part-Form	75	75	75	74	-	73	73	73	75	75	74	74	73	73	73	73
% Multidimensionality	100	100	0	0	-	100	0	0	100	100	0	0	100	100	0	0
Mean p-value	.70	.70	.67	.70	-	.72	.72	.71	.72	.72	.70	.71	.70	.70	.70	.70
s.d. p-value	.16	.16	.15	.14	-	.14	.14	.13	.16	.14	.15	.16	.17	.17	.16	.16
Mean abs. 2nd factor loading	.16	.16	.04	.04	-	.17	.05	.05	.16	.15	.05	.04	.16	.16	.04	.04
s.d. abs, 2nd factor loading	.07	.07	.03	.03	-	.07	.04	.04	.07	.07	.04	.03	.07	.07	.03	.03

Table 3

T statistics and p-values for the Stout Analyses of the Quarter Forms: 289 and 190

	Exam							
	289				190			
	1	2	3	4	1	2	3	4
% Multidimensionality	100	100	0	0	100	100	0	0
T statistic	3.23	2.47	-0.87	0.24	3.40	1.50	-0.29	0.91
p-value	.00	.01	.48	.40	.00	.07	.48	.18

Table 4

Concordance Rates (%) on Failure Decisions for the Quarter Forms
(For first % md form in each pair)

% md - % md (true)	Exam 288	mean	189	mean	289	mean	190	mean	Marginal Mean
100-100	43.80 60.83	52.32	- -	-	55.63 45.04	50.34	56.64 53.33	54.99	52.55
100-0	45.09 45.94 55.49 53.12	49.91	61.52 55.69	58.61	56.62 53.64 51.47 48.53	52.57	47.79 42.18 43.06 46.11	44.79	50.45
		52.10		55.40		52.91		50.10	52.23
0-100	56.27 49.87 60.56 50.42	54.28	51.46 52.91	52.19	51.66 58.01 48.80 54.52	53.25	61.13 58.49 47.19 54.79	55.40	54.01
0-0	51.47 54.37	52.92	52.93 60.11	56.52	60.73 60.54	60.64	49.43 43.23	46.33	54.10