DOCUMENT RESUME

ED 346 151 TM 018 492

AUTHOR Sykes, Robert C.; And Others
TITLE Dimensionality and DIF in a Licensure Examination.
PUB DATE Apr 92
NOTE 32p.; Paper presented at the Annual Meeting of the American Educational Research Association (San Francisco, CA, April 20-24, 1992).
PUB TYPE Reports - Research/Technical (143) -- Speeches/Conference Papers (150)

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS Factor Analysis; *Factor Structure; Health Personnel; Higher Education; *Item Bias; Item Response Theory; *Licensing Examinations (Professions); *Multidimensional Scaling; Nonparametric Statistics; Pass Fail Grading; *Test Construction; Test Format; Test Items
IDENTIFIERS Eigenvalues; *Parallel Test Forms; Rasch Model; Stouts Procedure

ABSTRACT

The sources of multidimensionality found in several different forms of a licensure examination were studied. The relationship between one source of multidimensionality, differential item functioning (DIF) (or factors producing DIF), and content characteristics was explored in an attempt to isolate aspects of training or curriculum that could account for the causes of multidimensionality in real data. A non-parametric approach for assessing unidimensionality developed by W. Stout (1987) was used to evaluate the dimensionality of several forms of a 300-item Rasch-based test used to license professionals in a health care profession. Quasi-random samples of 2,000 first-time candidates were selected for each of the 4 forms. The four forms were demonstrated to be multidimensional. The source of the multidimensionality could not be attributed to the presence of a large number of passage-like cases with associated multiple items. A second dimension was identified as the source of the multidimensionality through the magnitude of eigenvalue differences and the successful construction of part-forms made unidimensional by removal of items loading heavily on the second factor. An association between the second factor and items flagged for DIF was demonstrated. Items that loaded heavily on the second factor and were often flagged for DIF spanned content that involved knowledge and recall of physiological needs versus an evaluation/analysis of psychosocial needs. Six tables present study data, and there is a 16-item list of references. (SLD)

DIMENSIONALITY AND DIF IN A LICENSURE EXAMINATION

Robert C. Sykes
Kyoko Ito
Rachel Hols
Raymond T. Bradley

CTB Macmillan/McGraw-Hill

This paper was presented in April, 1992 at the Annual Meeting of the
American Educational Research Association
San Francisco, CA

For licensure examinations that are administered in multiple forms and calibrated using an IRT model, the issue of unidimensionality is important in two respects. First, since a licensure examination often generates a single score that is used to make decisions on whether candidates can enter the profession, it is imperative to ensure that different forms of the examination are equivalent to one another and consequently that pass-fail decisions made on the basis of different forms are consistent over forms.

A necessary condition to obtain equivalent forms for examinations that produce a single score is that all forms measure, to the same extent, one primary trait or ability to practice safely. Although use of a test plan in constructing the forms will contribute to the measurement of a common ability, the possibility still exists that the forms measure some other ability or abilities, in addition to the ability of interest, and therefore are not unidimensional. Multidimensional forms constructed on the basis of the same test plan may not be equivalent if uncontrolled and varying aspects of content significantly impact candidate performance. Thus, dimensionality of multiple forms of a licensure examination should be examined.

Second, unidimensionality is an important issue with licensure examinations that use a unidimensional IRT procedure for parameter estimation and the setting of pass-fail standards. Several researchers maintain that because testing usually requires more than one ability (e.g., reading skills in a math

test), real test data are inherently multidimensional (Harrison, 1986; Humphreys, 1986; Linn, Levine, Hastings, & Wardrop, 1981; Traub, 1983; Wang, 1988). Fitting a unidimensional IRT model to multidimensional data may result in model-misspecification error.

If the dimensionality of a form is evaluated and the form found not to be unidimensional, the sources of multidimensionality must be identified. While differential item functioning (dif) has attracted substantial attention as a potential source of bias, its relationship with multidimensionality has not yet been widely investigated with real data. Several researchers have, however, noted and demonstrated in simulations that multidimensionality may manifest itself as dif.

Lautenschlager & Park (1983) utilized the concept of multidimensionality in the generation of bias data: that is, by introducing a nuisance ability on which subject differences were confounded with differences on an ability of primary concern. As discussed by Wang (1988), an item can be either "multidimensional but unbiased" or "both multidimensional and biased". The former case may arise when the conditional distributions of the ability or abilities that the test was not designed to measure (i.e., nuisance abilities) are similar between two groups of examinees. On the other hand, when the two groups differ in the conditional distribution of the nuisance ability or abilities given the ability purportedly measured by the item, the item may be found to demonstrate dif.

Using simulated data, Oshima and Miller (1991) have shown that, irrespective of whether groups differ on the ability of interest, a small percentage of items that are multidimensional and biased can be correctly differentiated from a set of multidimensional but unbiased items. Also using simulated data, 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.

Given the relationship between multidimensionality and dif demonstrated in simulated data, one way to examine the possible causes of multidimensionality in real data would involve the following process. First, identify the items causing the test to be multidimensional. Second, examine these items by the Mantel-Haenszel method to see whether they manifest dif and finally depict them in terms of additional content characteristics. The depiction of content characteristics may have as one of its goals the exploration of common sources of multidimensionality and dif, such as differential training or educational effects (Traub, 1983). The identified sources of multidimensionality could then be controlled in a subsequent investigation of the practical impact of these factors on scores produced by unidimensional IRT models, such as the pass/fail classification decisions derived from them (Sykes, Ito, & Potter, 1992).

The purpose of this paper was to investigate the sources of multidimensionality found in a number of different forms of a licensure examination. The relationship between one source of multidimensionality - dif or factors producing dif - and content characteristics was explored in an attempt to isolate aspects of training or curriculum that could account for these phenomena.

Method

A nonparametric approach for assessing unidimensionality developed by Stout (1987) was used to evaluate the dimensionality of a number of forms of a 300-item Rasch-based test used to license professionals in a health care profession. The Stout approach assesses, through a significance test, the presence of a single dominant dimension. A complete examination is divided into two subtests: An assessment test which consists of items which maximally load or are judged a priori to maximally load on a second factor and a partitioning subtest consisting of all remaining items. Candidates are then divided into a number of homogenous groups on the basis of their scores on the partitioning subtest. The variance of candidate scores within each homogenous group is compared to the predicted unidimensional variance estimate for that group. These differences are then normalized and combined across groups to yield a statistic which can be assessed for the degree to which the average residual item covariance (controlling for each candidate's grouped level of performance) differs from zero. A test is "essentially"

unidimensional if, after accounting for item covariation due to the putative dominant dimension, residual item covariances are, on average, small in magnitude.

The Stout procedure was an especially appropriate procedure to evaluate the licensure examination because of the length of the forms and the "case-bound" nature of many of the items in the forms. There were no available factor analytic techniques that could provide a significance test of the number of factors underlying candidate performance for examinations as large as 300 items. Although the number of "case-bound" items (i.e., multiple items associated with a case of passage) was being reduced during the two-year period the licensure forms were administered, under the goal of their eventual complete elimination from the examination, each assessed form consisted of more than 50% case-bound items. The Stout procedure permitted a significance test of form dimensionality (i.e., the Stout T statistic referred to a standard normal distribution) that had been documented not to be susceptible to the contaminating effect of secondary dimensions affecting candidate performance on small sets of items, as could be predicted to occur for items associated with a single passage (Nandakumer, 1991).

The Stout procedure also incorporated an item tetrachoric factor analysis package that could be used to determine the items of the assessment test. A factor analysis of the items of each form, conducted on approximately 1000 candidates from each 2000 candidate sample, provided data such as eigenvalues and factor

loadings that permitted establishing the factor structure of any form that was found to be multidimensional by the Stout statistic.

For purposes of corroborating the Stout results, residual item correlations were computed for two of the forms after applying a nonlinear (cubic) factor analytic model (Etezodi, Amoli & McDonald, 1983). Distributions of residuals were compared across the two forms and with simulated unidimensional and multidimensional data reported by Hambleton and Rovinelli (1986).

Part-forms were also constructed to verify results obtained from the Stout analyses and evaluate hypotheses on possible sources of multidimensionality for one or more of the four forms. These part-forms were test-plan representative (i.e., proportionally meeting the test-plan content category quotas) and of an average difficulty that was similar to the four assessed forms.

Finally, content analyses were performed on sets of items identified to load most heavily on one or more of the two or three factors having the largest eigenvalues in Stout item tetrachoric factor analyses of one or more of the four forms. Included as part of these analyses were Mantel-Haenszel alpha (and transformed delta) statistics obtained from Mantel-Haenszel analyses of six ethnic groups for each of the evaluated items.

Sample

Quasi-random samples of 2000 first-time (i.e. the first time the candidates have taken the examination) U.S. educated candidates were selected for each of the four forms. First-time U.S. educated candidates serve as a large reference group for the licensing program. All classical and IRT (i.e. Rasch) examination and item statistics, with the exception of the Mantel-Haenszel statistics, are derived from samples selected from this reference group. The examination has repeatedly been demonstrated to be unspeeded for first-time U.S. educated candidates.

Of the four selected forms two were administered in the winter of 1989 and 1990 and are referred to as 189 and 190. The other two forms were from the second administration later in the calendar year. These two forms were administered in 1988 and 1989 and hence will be referred to as 288 and 289.

Results/Discussion

Assessment of Form Dimensionality

The 189, 289 and 190 forms were found not to be unidimensional ($T = 3.90$, $T = 3.61$, and $T = 2.73$ with $p < .001$, $p < .001$, and $p = .003$, respectively). The 288 form yielded a marginally insignificant Stout statistic ($T = 1.33$, $p = .092$). Nonlinear (cubic) factor analyses of the 288 and 189 forms resulted in mean residual correlations that were greater than two standard errors from the 0.0 predicted under unidimensionality.

The mean residual for the 288 form was .001 while the mean residual for the 189 form was .002, each mean based on 44,551 residual correlations. Although the means were in the direction of increased multidimensionality for the 189 form and both distributions were not normal by the Kolmogorov statatistic ($D = .006$ and $p < .01$ for both distributions (SAS, 1985)) the mean residuals were below mean residuals reported by Hambleton and Rovinelli (1986) for simulated two-dimensional data: .005 to .007.

The small size of the mean residuals relative to means obtained from simulated two-dimensional data and the large number of cases and case-bound items in the two examinations presented the possibility that mean residuals deviated from zero because of the presence of a large number of secondary dimensions associated with cases. As mentioned previously, Nandakumer (1991) demonstrated that results from the Stout procedure were not contaminated by the presence of secondary dimensions due to small sets of items associated with passages. However, she did not study examinations that had as many passages or cases as did the 300-item licensure forms. The 288 form had 60 cases, averaging 3.32 items per case while the 189 form had 61 cases, averaging 4.26 items per case.

In order to evaluate the possibility that the multidimensionality of the two forms was due to a large number of case dimensions, a half form was constructed from the full-length 189 form. Items were deleted from the 189 form, blind of extra-

test plan content, to produce a test-plan and difficulty representative half form that had only 16 cases, averaging 2.75 items per case. When evaluated by the Stout procedure, the half form was not unidimensional ($T = 3.03$, $p = .001$). Form multidimensionality that could not be attributed to secondary case dimensions was also indicated when nonlinear factor analysis residual correlations for the 288 and 189 forms were partitioned into between-case and within-case subsets. The mean within-case residual correlations were similar for the two forms (.012 and .010 for 288 and 189, respectively) while the mean between-case residual for the 189 form was two and a half times larger than that for 288 (.0020 vs .0008 respectively).

The eigenvalues produced by the item tetrachoric factor analyses of the four forms were examined to determine how many factors may be determining form dimensionality. The ten largest eigenvalues and differences between pairs of eigenvalues are presented in Table 1 for the four forms. Evaluation of these eigenvalue differences as well as those available from other analyses revealed that a difference between the second and third eigenvalue greater than .600 was always associated with a multidimensional Stout statistic. Conversely, a difference between the second and third eigenvalue 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.

The pattern of eigenvalue differences between the second and

third factors suggested the possibility that only one other factor, the second, was significantly impacting form dimensionality. In order to test this hypothesis items were deleted from each form that had large (absolute-valued) loadings on the second factor. Item deletion proceeded by deleting approximately equal numbers of items having positive loadings and items having negative loadings on the second factor. The remaining items were then verified to be test plan and difficulty representative and tested for unidimensionality using the Stout procedure. Because a previous attempt to create a unidimensional part-form by deleting a small number of items from one form (i.e., the 20 items constituting the Stout assessment subtest) did not produce a unidimensional part-form, item deletions began by deleting a minimum of 50 items having large second factor loadings (25 positive and 25 negative). Unidimensional part-forms could be created for all four forms by deleting between 100 and 143 items (all Ts' $\leq$ .64 and all ps' $\geq$ .25).

An attempt to create a test-plan and difficulty representative unidimensional part-form by deleting 77 items from one form that did not fit the Rasch model by the Wright and Panchapakesan's (1969) IRT fit statistic, evaluated at a p = .10 significance level, was not successful. (Examination items are typically screened for model fit on the basis of a smaller significance level). Hence, the multidimensionality of the four forms could be attributed to the presence of a second dimension whose effect could be attenuated by deleting items that loaded

heavily on the second factor but not by deleting items on the basis of model fit.

Characterization of the Second Dimension

For the purpose of characterizing the content forming the basis of the second dimension, the ten items which had the largest positive second factor loadings and the ten items that had the smallest (i.e. negative) second factor loadings were selected from each of the four forms. A content appraisal indicated that the content of the items on one pole of the second factor was similar to that of types of items often flagged for dif. Upon further analysis, a large number of the 40 items loading positively on the second factor were noted to have actually been flagged for dif against one or more of as many as six ethnic groups typically evaluated for minority group dif. These six ethnic focal groups are typically compared, using the Mantel-Haenszel procedure, to a majority white (reference) group and items flagged for dif against each of the six minority groups. The alpha cutscore of 1.81 had been previously determined to maximize the concordance of dif decisions with an IRT method of assessing dif (Sykes and Fitzpatrick, 1990). Majority group dif is currently not evaluated for this program.

The number of items that were flagged for dif against one or more of the ethnic groups out of each set of 10 items loading most positively and most negatively on the second factor are presented in Table 2. For comparison purposes, the number of

flagged items out of four sets of 10 items maximally loading on the unipolar first factor of each of the four forms and four paired sets of 10 items loading most positively and 10 items loading most negatively on the third factor for each form are also included. The three factors produced by the principal factor analytic solution are orthogonal to each other.

Proportionally more items loading extremely on the second factor were flagged for dif (28/80 = 35%) than those items loading heavily on the third factor (31%) or the first factor (10/40 = 25%). More noteworthy however, is the strong association between dif-flagged items and the poles of the second dimension. After reversing the polarity of the second and the third factor of the 190 form in order to match the direction of these factors for the other three forms, 26 out of 40 items loading most positively on the second factor (65%) were dif "associated". Of the remaining four out of five factor poles, the next strongest association of a pole with dif-flagged items is the 45% for the positive third factor pole.

The association between dif flagged items and one pole of the second factor is strikingly consistent across forms. A minimum of 50% of each of the four sets of 10 items loading most positively on the second factor are dif associated while no more than one item in only two of the four sets of 10 items having most negative second factor loadings were flagged for dif. The smallest difference between the number of flagged items across the four pairs of second factor poles, five for the 189 and 190

forms, is actually the largest difference obtained for the four pairs of third factor poles (6 - 1 = 5 for the 289 form). The marked pattern of dif flagged items associated with the positive second factor pole and not with the negative second factor pole suggests an association of dif or a factor or factors inducing dif with the content of the second factor.

This association is even more apparent when the number of flagging incidents or times that an ethnic group was flagged on items within the sets of 10 items is tallied. Presented in Table 3 the differences across the poles of the second factor are even more pronounced for each of the four forms, resulting in a ratio of 25 flagging incidence on the positive poles for every flagging incidence on the negative poles.

To facilitate the comparison of descriptive dif statistics on the assessed items, the Mantel-Haenszel alphas for each selected item for every available ethnic group - majority group comparison was transformed to a delta through the relationship:

$$_{mh} = -2.35 \times \ln \; _{mh}$$

The delta scale is symmetric around 0, with a negative delta signifying dif against the minority group and a positive delta dif against the majority group.

Mean deltas were then computed for each set of 10 items for each available ethnic group for three of the four forms: 288, 189, and 289[1]. These mean deltas are presented in Table 4 for

---

[1]Mean deltas were not available for the 190 form at the time this paper was submitted.

the positive poles of the first three factors and in Table 5 for the negative poles of the second and third factors. Of the three factors, only the second factor has universally negative delta means on one pole (i.e. the positive pole) and universally positive delta means on the other pole. Thus, the association of the positive second factor pole and dif or dif-associated factor(s) is also evident in these delta values.

The association of dif flagged items with the second factor prompted an assessment of the degree to which forms could be "purified" to be unidimensional by deleting items with extreme alphas. Because more items are typically flagged for dif against ethnic group four than against any other ethnic group, alphas for this minority group were used for determining item deletions. A correlation of -.53 ($p < .001$) between deltas for ethnic group 4 and second factor loadings across <u>all</u> the items in the 289 form substantiated a strong association between the two.

For each of the four forms, items with the most extreme alphas, both above <u>and</u> below 1.0, were deleted, the remaining items verified to be test plan and difficulty representative and subsequently tested for unidimensionality using the Stout procedure. Three of the four part-forms: 288, 289 and 190 were unidimensional ($T = 1.18$, $p = .12$; $T = .11$, $p = .46$; and $T = 1.47$, $p = .07$, respectively) by deleting the items having the most extreme alphas for ethnic group four (143, 155, and 150 items, respectively). The fourth part-form (189) remained multidimensional ($T = 3.09$, $p = .001$). Although one of the three

successfully purified part-forms was only marginally unidimensional, the effect on form dimensionality of deleting dif flagged items may be considered substantial in the light of the fact that the item deletion criterion was dif against only one ethnic group.

Content Characterization

Items from three of the four forms (288, 189, and 289) were evaluated by content experts. For each of the three paired sets of 10 items loading most positively and 10 items loading most negatively on the second factor, two content experts independently and blindly characterized the content in the following manner. The majority of items loading negatively required knowledge and recall of patients' physiological needs, although this characteristic was not noted for the corresponding items on the 288 form. (The 288 form was marginally unidimensional by the Stout statistic). The majority of items loading positively on the second factor for all three forms were noted to require analysis and evaluation, often of a psychosocial nature.

Thus the items loading most heavily on the second factor measured two types of professional expertise: knowledge recall of physiological needs and analysis/evaluation. The analysis/evaluation type of item that was frequently found in items located on the positive pole of the second factor often is associated with dif against minority groups (i.e., negative deltas). The particular knowledge/recall type of item often found among items on the negative pole of the second factor is frequently associated with positive deltas, implying a dif in favor of minority groups.

The fact that the items of interest spanned two different kinds of professional expertise explains the pattern of mean

deltas in Tables 4 and 5. In Table 4 the universally negative mean deltas on the ten items loading most positively on the second factor suggest that the dif or factor(s) inducing dif, impact, to varying degrees, all ethnic groups. Conversely in Table 5 the universally positive mean deltas on the negative pole of the second factor imply that the dif, or factor(s) inducing dif, favor on these items the performance of all ethnic groups relative to the majority white group. The type of broad effect manifested by the mean deltas on the second factor pole is not consistent with a type of dif manifested by culturally specific, or ethnic group specific aspects of content that would expectedly impact some ethnic groups and not others. For ethnic groups that are predominantly non-native, colloquialisms or idioms are examples of such content aspects.

Two aspects of professional expertise that require different abilities or skills, such as knowledge/recall of physiological needs versus evaluation/analysis of psychosocial needs, could account for such a broad effect if training of these abilities differed across different educational or training programs. Additionally, the ethnic groups would have to be more frequently exposed, relative to majority group candidates, to types of training programs that did not emphasize training of one type of ability, such as evaluation and analysis of psychosocial needs, while perhaps emphasizing training of the other ability or skill: knowledge or recall of psysiological needs.

Additional evidence for a contrasting ability effect,

perhaps attributable to different professional training, are the mean deltas for ethnic groups 4 and 5. These two groups are predominantly educated outside the United States in foreign schools with curricula that have been commonly noted to differ from those in U.S. schools. Specifically, the curricula offered by foreign schools emphasize the learning of knowledge of physiological needs required by health care professionals practicing in an institutional setting. They do not emphasize the training of analysis/evaluation of psychosocial needs or clinical skills required for practice in the noninstitutional settings which in this country are employing increasingly large numbers of professionals. These types of skills, often teaching and counseling in nature, are necessary for safe and effective practice and, in general, for facilitating a successful interaction with consumers of U.S. health care. Hence foreign curricula might be expected to compound the effect of candidates growing up in a foreign culture on candidate performance on examination questions that require consideration of U.S. social norms in analyzing and evaluating health care consumers.

Although the training of candidates from ethnic groups 4 and 5 is consistent with the appraised content characteristics of the second factor and their mean deltas, the fact that these candidates are predominantly foreign educated cannot explain the presence of a second factor in performance of samples of candidates who are U.S. educated. Furthermore the small proportion of the population of first-time U.S. educated

candidates constituted by candidates from ethnic groups 1, 2, 3, and 5 means that the presence of the second factor cannot be attributed to the training of candidates from these ethnic groups within the selected samples. The existence of a second factor among first-time U.S. educated candidates may be explained however, by the presence of training programs in U.S. schools that emphasize the training of these two broad types of abilities.

The hypothesis that a second factor, associated with the performance of candidates educated in the U.S. and manifested in the performance of candidates educated outside the U.S. through a broad type of dif, would be supported if the dimensionality of the performance of U.S. educated candidates differed across U.S. educational programs that also differed in their emphasis of the training of the two types of skills. In order to evaluate this hypothesis, samples of 2000 candidates from each of the three different types of educational programs offered in the U.S. were selected, where numbers permitted, from the 288, 189, and 289 administrations. Stout analyses were performed on samples available for all three forms for educational program 1, all three forms for program 2, and the 288 and 289 forms for educational program 3.

The results presented in Table 6 verify that the dimensionality of candidate performance varies over educational programs. For educational programs 1 and 2, which together train more than 90% of the first-time U.S. educated candidates,

candidate performance is multidimensional with only one exception: the 289 form/sample for the second program was unidimensional ($T = -2.97$, $p = .48$). These two programs generally expose their students to a wide spectrum of clinical training. The third type of educational program historically has often offered its students clinical training more geared to an institutional setting. Students educated in the third program, like foreign educated candidates, consequently may have less exposure to the other broad type of non-institutional work environments. The dimensionality of the performance of candidates for the two available program form/samples, 288 and 289, was unidimensional though in one case marginally so ($T = 1.54$, $p = .06$ and $T = 1.16$, $p = .12$ respectively).

It should be noted that while the dimensionality of candidate performance varies over educational program, these differences do not produce significant differences in passing rates across programs. Candidates trained in the third educational program pass at a rate that is very similar to the rates for the other two programs. In fact, the passing rate for candidates from the third program is often, though slightly, the highest passing rate for the three programs.

Conclusions

Four forms of a licensure examination were demonstrated to be multidimensional using the Stout procedure for assessing "essential" unidimensionality. The source of the multidimensionality could not be attributed to the presence of a large number of passage-like cases with associated multiple items. A second dimension was identified to be the source of the multidimensionality through the magnitude of eigenvalue differences and the successful construction of part-forms made unidimensional by removal of items loading heavily on the second factor. Candidate performance that is demonstrated to be multidimensional might have a practical impact on not only examinations that generate a score based on an IRT model explicitly assuming unidimensionality but any examination that produces a single score. The practical impact of the second factor on test scores produced from an IRT-based model was investigated in additional work (Sykes, Ito, and Potter, 1992) and no practical effect was found.

An association between the second factor and items flagged for dif was demonstrated. Items that loaded heavily on the second factor and were often flagged for dif spanned content that involved knowledge and recall of physiological needs versus an evaluation/analysis of psychosocial needs. These two types of abilities or skills may be emphasized to a different degree in the professional training of foreign educated candidates.

It may be possible that this is also the case in different training programs offered to U.S. educated candidates, though to a much lesser extent than for foreign educated candidates. The dimensionality of the performance of candidates educated in the

three types of U.S. programs differed in dimensionality across programs that also differed in the type of clinical training offered. The differences in the dimensionality of candidate performance across programs, however, do not apparently impact passing rates which are very similar across the three programs. Additional work is needed to verify that the nature of the dimensionality of foreign educated candidate performance is similar to that obtained for a U.S. program.

References

Ackerman, T. A. (1988, April). An explanation of differential item functioning from a multidimensional perspective. Paper presented at the Annual Meeting of the American Educational Research Association, New Orleans.

Etezadi-Amoli, J. & McDonald, R.P. (1983). A second generation nonlinear factor analysis. Psychometrika, 48. 315-342

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

Hambleton, R.K & Rovinelli R.J. (1986). Assessing the dimensionality of a set of test items. Applied Psychological Measurement, 10, 287-302.

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

Lautenschlager, G. J., & Park, D. G. (1988). IRT item bias detection procedures: Issues of model misspecification, robustness, and parameter linking. Applied Psychological Measurement, 12, 365-376.

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.

Nandakumar R. (1991). Traditional dimensionality versus essential dimensionality. Journal of Educational

Measurement, 28, 99-117.

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.

SAS Institute. (1985). SAS User's Guide: Basics (1985 ed.)- Cary, NC: Author.

Stout, W. (1987). A non-parametric approach for assessing latest trait unidimensionality. Psychometrika, 52, 589-617.

Sykes, R.C. & Fitzpatrick, A.R. (1990). Establishing a Mantel-Haenszel alpha cutscore through a multiple method procedure. Paper presented at the Annual Meeting of the American Educational Research Association, Boston, MA.

Sykes, R.C., Ito, K., and Potter, R. (1992). Assessing the impact of multidimensionality on the classification decisions of an IRT-based licensure examination. Paper presented at the Annual Meeting of the National Council on Measurement in Education, San Francisco, CA.

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.

Wright, B.D. & Panchapakesan, N. (1969). A procedure for sample-free item analysis. Educational and Psychological Measurement, 29, 23-48.

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)	

Table 2

Number of Items of the Ten Highest Positive and Negative Loadings on the First Three Factors That were Flagged for Dif. (Alpha >= 1.81)

	Examination				
Loadings	288	189	289[s]	190[s,a]	Total
			1st Factor		
Positive	1	3	4	2	10
			2nd Factor		
Positive	8	6	7	5	26
Negative	1	1	0	0	2
			3rd Factor		
Positive	3	3	6	5	17
Negative	3	2	1	2	8
Total	16	15	18	14	63

[s] 289 and 190 had more ethnic categories (6) than 288 or 189 (4).

[a] The polarity of the second and third factors was reversed for 190 to match the direction of these factors for the other exams.

Table 3

Number of Flags/Incidents (Alpha >= 1.81) on the Ten Items With the Highest Positive and Negative Loadings on the First Three Factors

	Examination				
Loadings	288	189	289[$]	190[$@]	Total
			1st Factor		
Positive	1	4	10	6	21
			2nd Factor		
Positive	12	12	14	12	50
Negative	1	1	0	0	2
			3rd Factor		
Positive	3	5	10	9	27
Negative	6	4	1	4	15
Total	23	26	35	31	115

[$] 289 and 190 had more ethnic categories (6) than 288 or 190 (4).

[@] The polarity of the second and third factors was reversed for 290 to match the directions of these factors for the other exams.

Table 4

Mean Deltas by Ethnic Group
for the Ten Items With the Highest Positive
Loadings on the First Three Factors

Exam	Ethnic Group 1	2	3	4	5	6	Mean[a]
				1st Factor			
288	.18	-	-	.70	.21	.19	.32
189	-.26	-	-	.37	-.01	-.01	.02
289	.11	.10	-.32	-.42	-.58	-.03	-.19
				2nd Factor			
288	-.85	-	-	-2.09	-1.53	-.85	-1.33
189	-1.06	-	-	-.90	-1.00	-.68	-.91
289	-1.00	-.82	-.68	-1.89	-1.29	-.77	-1.08
				3rd Factor			
288	-.19	-	-	-.32	-.24	-.32	-.27
189	.54	-	-	-.43	-.65	-.26	-.47
289	-.38	.14	-.78	-.89	-1.03	-.41	-.56
Mean[a]	-.44	-.19	-.59	-.65	-.68	-.35	-.51

[a] Mean of individual deltas and not the mean of means.

Table 5

Mean Deltas by Ethnic Group
for the Ten Items With the Highest Negative[§]
Loadings on the Second and Third Factors

Exam	Ethnic Group 1	2	3	4	5	6	Mean[a]
				2nd Factor			
288	.76	-	-	1.80	1.31	.44	1.08
189	.97	-	-	2.51	2.08	.70	1.57
789	.85	.23	.70	1.94	1.43	.65	.97
				3rd Factor			
288	-.54	-	-	-.62	-.76	-.17	-.52
189	.24	-	-	1.28	.54	.01	.52
289	.59	-.09	1.01	1.13	1.46	.41	.75
Mean[a]	.48	.07	.86	1.34	1.01	.34	.75

§ The first factor was unipolar.
[a] Mean of individual deltas and not the mean of means.

Table 6

First 10 Eigenvalues from the Linear Factor Analyses of the Educational Programs Assessed for Dimensionality

Program Samples

Educational Program: 1

288		189		289	
Eigenvalue	Difference	Eigenvalue	Difference	Eigenvalue	Difference
16.568	12.345	17.599	12.218	17.442	12.606
4.223	0.825	5.382	1.740	4.836	1.030
3.398	0.226	3.642	0.481	3.806	0.259
3.171	0.023	3.161	0.074	3.547	0.235
3.148	0.243	3.087	0.222	3.313	0.033
2.905	0.091	2.865	0.018	3.280	0.189
2.814	0.042	2.848	0.082	3.091	0.087
2.772	0.084	2.766	0.092	3.004	0.107
2.689	0.036	2.674	0.086	2.896	0.079
2.652	0.091	2.588	0.043	2.817	0.062
T = 3.61 sign. (p < .01)		T = 5.64 sign. (p < .01)		T = 2.34 sign. (P = .01)	

Educational Program: 2

18.636	14.209	17.620	12.690	19.259	14.816
4.427	0.605	4.931	0.832	4.444	0.382
3.821	0.111	4.099	0.159	4.062	0.591
3.711	0.605	3.940	0.759	3.671	0.361
3.106	0.082	3.181	0.191	3.110	0.063
3.024	0.061	2.991	0.039	3.047	0.165
2.963	0.110	2.952	0.181	2.882	0.067
2.853	0.083	2.772	0.049	2.815	0.075
2.769	0.122	2.722	0.043	2.740	0.097
2.647	0.010	2.679	0.015	2.643	0.034
T = 4.66 sign. (p < .01)		T = 3.23 sign. (p < .01)		T = -2.97 n.s. (P = .48)	

Educational Program: 3

16.023	11.606			15.238	10.983
4.417	0.447			4.255	0.298
3.970	0.525			3.957	0.576
3.445	0.093			3.381	0.083
3.351	0.125			3.298	0.170
3.226	0.198			3.128	0.052
3.029	0.198			3.075	0.120
2.831	0.017			2.955	0.033
2.814	0.046			2.923	0.032
2.767	0.033			2.891	0.138
T = 1.54 n.s. (p = .06)				T = 1.16 n.s. (p = .12)	