

PSYCHOMETRIC CHALLENGES IN DEVELOPING A COLLEGE ADMISSION TEST FOR JORDAN

MOUSA ALNABHAN
Mutah University, Jordan

MICHAEL HARWELL
University of Minnesota, USA

In 1998, the Jordanian Council of Higher Education authorized the construction of a standardized aptitude test that would be used to assist colleges and universities in admissions decisions. In this paper we report the results of a study in which it was examined whether test items were operating as desired and path analyses that explored predictors of student performance for a highly selective sample of Jordanian students. Item analyses indicated that 30% of the items showed inadequate discrimination or inappropriate difficulty levels, and an additional 19% of the items showed evidence of differential item functioning attributable to sex. The path analyses indicated that the strongest predictors of performance emerged for female students and included parental educational level and whether students attended a government-sponsored school or a private school. For males, the same predictive relationships were negligible.

Keywords: college admission, aptitude test, students, performance, Jordan.

The use of standardized tests in college and university admission decisions is well established in some countries but less so in others, including Jordan. In 1998, the Jordanian Council of Higher Education authorized construction of a standardized aptitude test that would be used to assist colleges and universities in admissions decisions. The rationale for the Council's decision lies in Jordan's dramatic population growth and the anticipated need for a skilled and well-educated labor

Mousa Alnabhan, Associate Professor in Education Research Methodology, Mutah University, Jordan. Michael Harwell, Professor of Educational Psychology, University of Minnesota, USA.

Appreciation is due to anonymous reviewers.

Please address correspondence and reprint requests to: Mousa Alnabhan, Department of Psychology, United Arab Emirates University, P.O. Box 17771, Al-Ain, United Arab Emirates. Phone (+971) 50-563-9865; Fax: (+971) 3-767-1705; Email: mousa_nabhan1@excite.com

force in the 21st century. The availability of an admissions test with strong psychometric properties with content closely linked to the skills needed to succeed in college can play an important role in this process. In this study we investigated the psychometric characteristics of the test using data from a pilot study.

Although empirical investigations of new tests are fairly standard, two factors complicated this work: (1) The sample of Jordanian students used in the pilot study represented the top 5-15% of academically talented 10th and 11th grade students in Jordan, as nominated by their secondary school, raising the possibility of restricted variation in the data; (2) The educational advantages that often accrue to males in Jordanian society made it likely that strong sex differences existed in test performance. Both of these factors could pose significant psychometric challenges to the test development process.

The authors investigated also the role of other variables that may affect test scores. One was school authority, which essentially reflects the organization sponsoring the school attended by students (government versus private). In Jordan, private schools are beyond the reach of many citizens because these schools have high tuition rates and fees, and most children from poor and middle-class backgrounds attend schools sponsored by the Jordanian government. As a result, school authority provides a rough measure of a student's socioeconomic status, and it was hypothesized that this variable would affect test performance because of the large economic and social differences between the two types of schools.

In addition, the relationship between student achievement and some family-based factors was investigated. In earlier work a negative relationship was found between the achievement of Jordanian students and family size, and a positive relationship between achievement and parental educational level (Abu-Lebdeh & Ahlawat, 1997; Alnabhan, 1997). These findings led the authors to hypothesize that students from larger families would be less likely to perform well because, in Jordanian society, large families are often associated with lower levels of parental education. Parental education was also expected to positively affect test performance. For example, students from better-educated families may outperform those from families in which parents do not have much education, although this effect may be mediated by the academic success of students.

Our purpose in this study was to examine the psychometric properties of the test to determine if they were adequate for test development to be continued; other researchers and practitioners were responsible for ensuring that the resulting test closely matched the skill domains to be assessed. The authors were particularly interested in examining the effects of restricted variation in the test response data and the magnitude of sex differences. Secondary questions of interest included learning if school authority, father's education, mother's education, or family size affected test performance. Currently, no empirical investigation of the test has been undertaken, and these results will help to guide subsequent revisions.

The construction of the test began in 1999 with a national panel consisting of content experts in Math, Science, English, Arabic, and Measurement and Statistics. These experts were charged with writing items that assessed specified domains, for example, verbal skills. With the construction of 100 dichotomously-scored items, each with five options, the panel felt that the test was ready to be piloted. Like most college entrance examinations, the test is intended to be norm-referenced and is described as an *aptitude test*.

To pilot the test a sample of 452 Jordanian students sat the examination in April 1999. As noted above, these students were among the top 5-15% of academically talented 10th and 11th graders in Jordan – although widespread use of the test should result in a greater range of academically talented students sitting the examination, perhaps as high as the top one-third. Still, the sample should be representative of the population of 10th and 11th graders who are interested in attending college and who are most likely to be admitted, even though it was a convenience sample. Most of the students in the sample (68.1%) attended a secondary school sponsored by the government, a value close to that for secondary school students in Jordan as a whole. Similar comments hold for variables reflecting father's education (48.9% had no formal schooling), mother's education (75.7% had no formal schooling), family size (ranging from 1 to 6 children), and sex of students (66.7% males).

Students were expected to be highly motivated because those who score well may have the opportunity to study abroad by having their college work sponsored by the World United Colleges (WUC), which are special colleges for gifted students from around the world. However, it is possible that males approached the test with greater motivation and enthusiasm than did females. In Jordanian society, study abroad tends to be discouraged for females, whose roles are often viewed as being nurturers and mothers whose place is in the home. This possibility heightened the concerns of the authors over sex differences in test performance.

METHOD

Various psychometric analyses were performed that were guided by four conditions set down by the expert panel as minimally necessary for adoption of the test:

(1) Responses across examinees should show adequate variability at the test and item levels; (2) Items should show psychometric characteristics that are consistent with the purpose of the test; (3) Test performance should not depend on the form of the test with which a student was presented; (4) Items should not behave differently for male and female examinees.

The authors began by examining test and item level variability. Then the behavior of each test item was studied to determine if it was operating as desired using item response theory (IRT). Next, the possibility of an item presentation effect was

investigated by testing whether student performance differed as a function of which form of the test they received. Fourth, the data were examined for evidence of differential item functioning (DIF) for male and female students. Path analysis was used to study the effects of other variables believed to affect test performance.

RESULTS

COMPLIANCE OF TEST WITH THE FOUR CONDITIONS

Condition 1: Responses across examinees should show adequate variability at the test and item levels

Ensuring adequate variability in responses is particularly important because of the highly selective nature of the sample. The authors began by creating a total score for each student by summing the number of correct responses. Since each of the 100 items was dichotomously scored (correct, incorrect), the possible total score range was 0-100. There were missing data for 15 students, who failed to respond to 10-15 of the items; these cases were deleted from the original sample, leaving a final sample of 437.

Although the missing data may not be missing completely at random (Schafer, 1997), which would ensure that no bias accrues because of the deletion, the small number of deleted cases makes it likely that the magnitude of any bias is small and random.

To examine test-level variability, the total correct scores were plotted (see Figure 1). Results indicated that the distribution was unimodal and almost symmetric ($M = 59.7$, $Mdn = 59$, skewness = $-.09$). The SD of 11.29 suggests that there was adequate spread in the total correct scores.

To examine item variability, the authors began by computing the percentage of responses across the five options for each item. Averaged across 100 items, these values were 20.27% (26.19), 16.13% (21.6), 19.51% (23.21), 18.79% (21.26), and 24.21% (28.96); values in parentheses are SD s. These percentages provide global evidence of variability in student responses because the average percentages of students selecting one of the five options were similar. For example, on average, 20.27% of the students responded to the first option for the 100 items, and 16.13% to the second option. The fact that the SD s are larger than the means also provides evidence of variability in responses.

The authors computed also the percentage of students answering each item correctly. Items that are too easy or too difficult would show percentages near one or zero, respectively. For 8 of the 100 items, more than 90% of the students answered these items correctly; however, none of the items had less than 10% of the students answering correctly. The eight items with percentage-correct values above 90% are likely to be inappropriate for a norm-referenced test and were flagged for special attention in subsequent analyses.

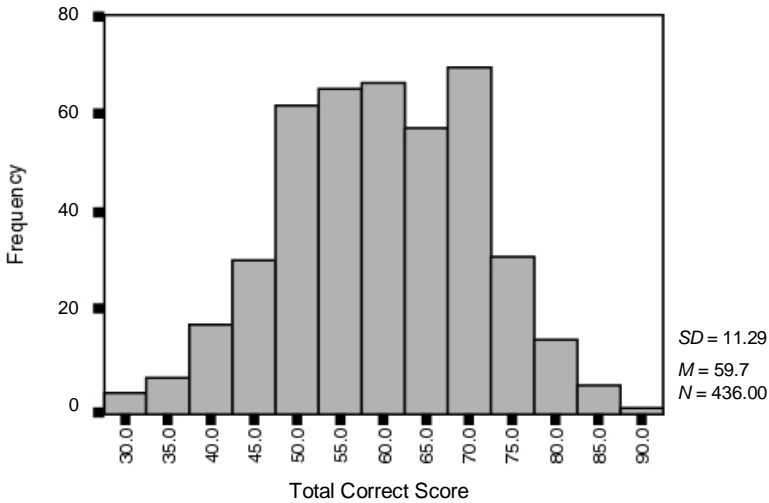


Figure 1. Frequency of total correct scores.

In sum, descriptive analyses provided no evidence of variability, restriction problems at the test level, and the authors' concern over restricted variability appears to be unfounded at the test level. This suggests that data analyses using total correct scores should not be affected by variability problems. However, there was evidence of some restricted variability at the item level as reflected in the item difficulties, and analyses examining the operating characteristics of items will have to attend to potential variability problems.

Condition 2: Items should show psychometric properties that are consistent with the purpose of the test

Because the test is intended to be norm-referenced, comparing student performances is important. Summing the number of correct answers to create a total score provides an estimate of each student's true proficiency and is consistent with classical test theory (CTT). However, CTT has several deficiencies (Hambleton & Swaminathan, 1985), including its failure to incorporate specific properties of each item into the estimation of each student's proficiency and the fact that total test scores under CTT represent an ordinal scale of measurement (Embretson, 1996; Fischer, 1995). Item response theory (IRT) does not possess the deficiencies of CTT and was used to study the psychometric properties of each item and to estimate each student's proficiency. Because guessing on multiple choice items is not common among Jordanian students, two-parameter IRT logistic models were fitted to the responses of each item using the BILOG for

Windows program (Mislevy & Bock, 1997). These results allowed the authors to study how each item was operating.

Before continuing, it was noted that IRT models require the data to be unidimensional or essentially unidimensional. Factor-analytic results indicated the presence of a dominant factor and several minor factors (eigenvalues between 1.0 and 2.0). This is not surprising, because the test was not developed with the intent of having subtests representing different domains (factors) of proficiency, despite the fact that the items are intended to assess skills in Math, English, etc. Because of this, the factor structure is somewhat muddled and any revisions of the test must take the notion of factor structure into account. Still, the pattern of small factors suggests that the data are essentially unidimensional, meaning that IRT models can be applied cautiously to these data.

In using the two-parameter logistic IRT model, it was *a priori* decided that, conditional on adequate model-data fit, items with estimated difficulty parameters within ± 3 standard deviations of zero (assuming difficulty parameters follow a standard-normal distribution) would be considered to be operating acceptably; items with difficulty parameters outside this range would be judged to be too difficult or too easy. Similarly, estimated discrimination parameters greater than .50 would be considered to provide adequate discrimination between students showing more or less proficiency – which is consistent with other cutoffs for which discrimination parameters (e.g., Muthen, Kao, & Burstein, 1991), particularly in the early stages of test development.

Overall, the average estimated difficulty parameter was .36 ($SD = .37$), and the average estimated discrimination parameter was .75 ($SD = .28$). In examining the items, it was found that three had estimated difficulty parameters beyond ± 3 standard deviations, all of which were negative, and 15 had estimated discrimination parameters under .50. On the whole, the test was somewhat easier than desired and certainly less discriminating than the expert panel had hoped. These findings, however, may be at least partly attributable to the highly selective nature of the student sample. For example, the average difficulty value is probably the result of using a sample of academically talented students.

Next, the model-data fit for each item was assessed using a chi-square fit test. If model-data fit is adequate then interpretations of parameter estimates are more credible; if model-data fit is inadequate then credible interpretations of parameter estimates are usually not possible. Using $\alpha = .05$, eight of the chi-square fit tests were statistically significant, meaning that the model-data fit was unacceptable. Three of the eight items showing inadequate model-data fit also had unacceptably large difficulty parameters. The eight items were flagged for consideration later. On the whole, though, the effects of restricted variation at the item level appeared to be modest, and the estimated difficulty and discrimination parameters appear to be reasonable estimates of the operating characteristics of most items.

Condition 3: Test performance should not depend on the form of the test with which a student was presented

Each student completed all 100 items, with items being presented in one of five arrangements (i.e., test forms). Plots of total scores for each of the five forms (based on 100 items) were similar to Figure 1 and are not presented. Two analyses were done to study item presentation effects. First, a one-way, fixed effects ANOVA was performed with Test Form as a between-subjects factor with five levels, and estimated proficiency (from BILOG) as the dependent variable. The results indicated no difference in average performance across test forms at $\alpha = .05$ ($F = 2.47$, $p = .12$); similar findings occur if total correct score serves as the dependent variable since high scores may be achieved by subjects of high proficiency. Because of the possibility that the test forms could show the same means but different variances, and such variance differences may imply an item presentation effect, the variances of the forms were also compared. Using the Levene test, it was found that the variances were not statistically different at $\alpha = .05$ ($p = .83$). It was also found that the reliability of the test was not affected by which form students were given, with internal consistencies ranging from .84-.87 for the five forms. These results suggest that the order of presentation of items to students had no effect on test performance.

Combining the above results, eight items had correct response rates exceeding 90%, eight showed poor model-data fit, and 18 showed difficulty and/or discrimination parameters not satisfying the a priori criteria. Collectively, $8 + 8 + 18 = 34$ items showed undesirable psychometric properties, but four items showed multiple problems (e.g., poor model-data fit and an estimated discrimination parameter less than .50). As a result, 30 of the original 100 items were flagged and information about their poor performance was transmitted to the expert panel who will decide whether they should be modified or deleted altogether in a revision of the test; the 70-item reliability was .83. Subsequent analyses were based on standardized proficiency estimates (mean = 0, $SD = 1$) from BILOG for the 70 items.

Condition 4: Items will not behave differently for male and female examinees

An examination of proficiency estimates for the 291 males and 145 females showed that there were moderate differences ($Y_{\text{Males}} = .087$, $Y_{\text{Females}} = -.116$; $SD_{\text{Males}} = .82$, $SD_{\text{Females}} = .79$), meaning that, on average, males outperformed females by approximately $|.087 - (-.116)| = .20$ standard deviations. The possibility of DIF due to sex was a concern throughout the test development process. To explore this, the authors fitted logistic regression models to the data for each of the 70 items following the methods used by Swaminathan and Rogers (1990). The presence of uniform and/or non-uniform DIF earned an item a DIF classification.

The logistic regression program in SPSS for Windows (1997) was used to perform the analyses. The variable Y_{ij} represented the dichotomous response of the i th

student on the j th item ($1 = \text{correct}$, $0 = \text{incorrect}$) and served as the dependent variable, and X_{i1} , served as a predictor variable and represented each student's estimated proficiency. The logistic regression model also contained a predictor X_{i2} that represented the sex of each student ($1 = \text{female}$, $2 = \text{male}$), and a predictor X_{i3} that represented the proficiency \times sex interaction. The focus was on uniform DIF for sex (with proficiency held constant) and non-uniform DIF. If the estimated slope for sex was significant (with proficiency held constant) then the item showed uniform DIF, meaning that the probability of answering correctly was uniformly higher for males or females. The contribution of interactions was assessed after entering proficiency and sex; a significant interaction signals the presence of non-uniform DIF, so that the relationship between Y_{ij} and proficiency depends on sex.

Each of the logistic regression analyses provided a test of the overall statistical significance of the model using the Wald statistic and tests for individual predictors (assuming that the overall Wald statistic was statistically significant at $\alpha = .05$). The statistically significant slopes are reported in Table 1 in exponentiated form and show that 16 items produced uniform DIF for sex. For example, for item15, the exponentiated slope for sex was 3.14, meaning that, with proficiency held constant, males were about three times more likely to answer this item correctly than were females, suggesting considerable DIF. Most of the exponentiated slopes for uniform DIF are substantial and positive, suggesting that these items should be carefully reviewed for an explanation of why males often had a greater probability of responding correctly.

TABLE 1: STATISTICALLY SIGNIFICANT SLOPES

Uniform	Item	2	15	19	21	36	50	59	69	85	88	92	93	94	96
DIF	Exp B (sex)	.64	3.14	2.36	.50	.66	.63	.52	1.1	1.7	2.3	1.7	1.8	1.5	1.7
Non-Uniform	Item	15	16	36	47	80									
DIF	Exp B (sex*pr)	2.63	1.87	1.74	2.5	2.6									

* Pr: Proficiency
 All reported slopes were significant at $\alpha = .05$.

Five items showed significant non-uniform DIF (two items produced both uniform and non-uniform DIF). Altogether, 19 of the 70 items showed DIF, and most of these showed large exponentiated slopes. These 19 items were flagged for further study by the test developers and content experts to determine if they show bias. Items showing bias will be modified or deleted in a revision of the test.

Collectively, this investigation of the psychometric characteristics of the test suggests that item responses do show some restriction of variability. There is also evidence of strong sex differences in test performance, with males having a higher probability of correctly answering several of the items as well as showing higher total correct or proficiency scores on average. It is likely that the pre-eminent role of males in Jordanian society, reflected in the greater opportunities and higher educational level available to them compared to females, explains these findings. A significant revision of the test is needed that takes these findings into account.

PATH ANALYSES

Path analyses were used to answer questions about the effects of various variables on test performance. Specifically, the effects of school authority (government-sponsored or private), parental education, and family size on student performance, and the extent to which these relationships might be mediated by the student's own academic success were explored. Path analysis allowed the authors to test model-data fit, which is crucial to the credibility of the statistical inferences, and to estimate and test indirect effects. All variables were treated as manifest; latent variable models were not used because of the lack of indicator variables as well as the relatively small sample size. In all cases, student proficiencies estimated by BILOG for the 70 items remaining after earlier analyses served as the dependent variable. The items showing evidence of DIF were included in the proficiency estimates because it is yet to be demonstrated that they are biased.

The initial path model was developed based on results for Jordanian samples in Abu-Lebdeh and Ahlawat (1997) and Alnabhan (1997) and is presented in Figure 2. This figure posits that school authority (SchAuth), father's education (FathEd), mother's education (MothEd), family size (FamSize), and grade point average in ninth grade (GPA9th) have direct effects on student proficiency (Profic). The authors examined also whether the effects of FathEd, MothEd, and SchAuth on Profic were mediated by their children's academic performance, i.e., whether a student's academic performance can strengthen or weaken the effect of these predictors on Profic. For example, they posited that FathEd would have both a direct effect on Profic as well as an indirect effect (through GPA9th). Suppose that a student comes from a family in which parents have little or no formal education and that this is weakly associated with lower test scores. The effect of strong academic credentials, reflected in high GPA8th values, may be to mediate the effect of parental educational level.

Two pieces of evidence of a student's academic credentials were used by including GPA8th as a predictor of GPA9th and, indirectly, of Profic. Because of the traditional role of males as providers and females as mothers and nurturers in Jordanian society, the authors fitted the path model in Figure 2 to male ($N = 291$) and female ($N = 145$) samples separately. The sample sizes are small but are simi-

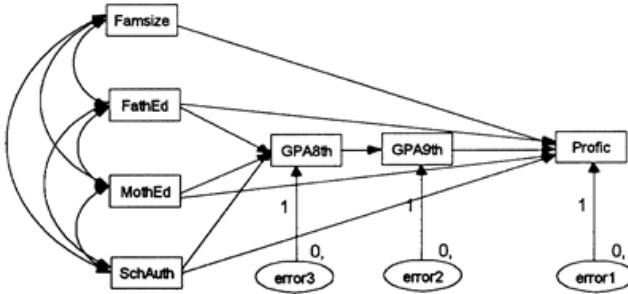


Figure 2. Initial path model.

lar to those reported by Hatcher (1996), Mueller (1996), and Kacmar and Valle (1997) and should be large enough to enable effects of interest to be detected.

The path analyses were performed using the LISREL (Joreskog & Sorbom, 1993) and AMOS (Arbuckle, 1997) computer programs. The authors began by fitting the model in Figure 2 to the samples of male and female students. Each of the fitted models showed satisfactory overall model-data fit. The chi-square tests were not statistically significant at $\alpha = .05$ for the two models, all of the standardized model residuals were less than ± 2 as recommended by Hayduk (1987), and the NFI and CFI fit indices exceeded the .90 cutoff as recommended by Hu and Bentler (1995). The fitted model for males is presented in Figure 3 and for females in Figure 4. The values on the paths are the standardized path coefficients.

The magnitude of these coefficients indicates that most of the relationships in these models are modest. For example, the largest (and only) statistically significant path coefficient for males was the effect of GPA8th on GPA9th (.70), meaning that, with the other relationships held constant, each one standard deviation increase in GPA8th was associated with a .70 standard deviation increase in GPA9th. The nonsignificant path from SchAuth to Profic means that the nature of the secondary school attended by these students (government-sponsored versus private) had no effect on their test performance. Also, GPA8th did not mediate the effects of FathEd and MothEd on Profic. However, the indirect effect of SchAuth on Profic (not reported in Figure 3) was statistically significant (-.11), suggesting that students with higher GPA8th values tend to have higher proficiency scores even if they attend a government-sponsored school. Still, almost all of the relationships in Figure 3 are either not statistically significant, or weak.

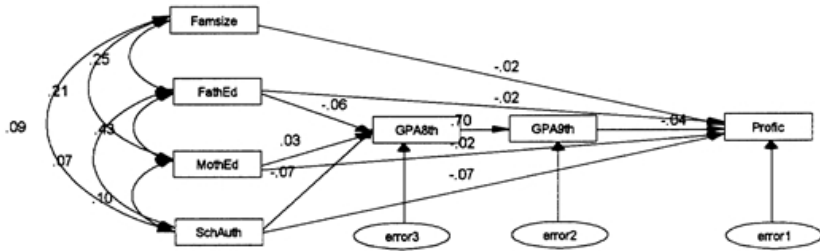


Figure 3. Fitted model for males.

A different pattern appears in Figure 4. Several effects emerged, including a strong direct effect of FathEd on Profic (-.29), meaning that, with the other predictors held constant, female students whose fathers had some formal education tended to perform more poorly than female students whose fathers had no formal education. This finding is slightly less puzzling in light of the direct (positive) effect of MothEd on Profic (.17), which indicates that female students from families in which the mother has some formal education tend to show greater proficiency than do females from families in which the mother has no education. It might be conjectured that part of this effect is attributable to more formal education resulting in some parents having less contact with their children because it is common in Jordan for jobs to require that parents (particularly fathers) be away from home frequently. For these individuals, less contact may translate to their having less effect on their child’s achievement.

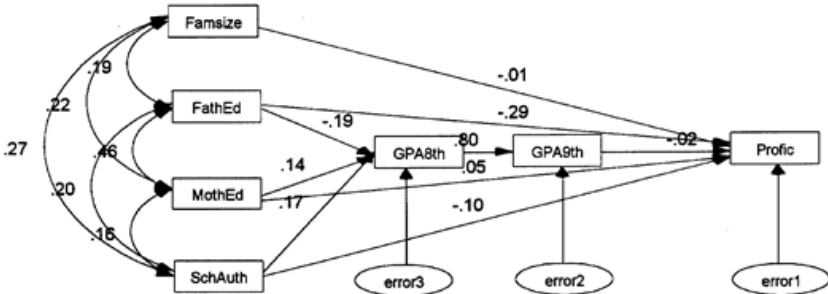


Figure 4. Fitted model for females.

An even more likely explanation for these findings is that the strong mother-daughter bonds in Jordanian society manifest themselves through educated mothers pushing their daughters academically, more so than educated fathers, and that the effect is enhanced by attending a private school. This conclusion is supported by the average proficiency scores for female students from families in which both parents had some formal education ($Y = .087, N = 96$) and families in which only the mother had some formal education ($Y = .113, N = 127$). Females from families

in which the mother had a formal education and who attended private school scored significantly higher ($Y = .12, N = 84$) than did those from families in which the mother had no formal education ($Y = .031, N = 213$); attending a government-sponsored school appeared to reduce the effect of MothEd ($Y_{\text{mother no educ}} = .006, N = 22$; $Y_{\text{mother educ}} = -.071, N = 117$). A similar explanation applies to the FathEd to GPA8th effect (-.19).

Females showed a somewhat different pattern of indirect effects from that shown by males. Like males, SchAuth had a significant indirect effect on Profic (-.11) and a nonsignificant indirect effect through MothEd. However, FathEd had a moderately strong indirect effect on Profic (-.17), suggesting that the effect of higher GPA8th values is to mediate the effect of FathEd on Profic.

In sum, male and female students produced different patterns for three predictive relationships: FathEd Profic, FathEd GPA8th, and MothEd Profic. The fitted male and female path models were compared by constraining the path coefficients of interest and testing whether the constrained models differed. In this way the authors learned that these predictive relationships differed significantly for males and females. For example, the FathEd Profic coefficients of -.02 and -.29 for males and females, respectively, differed. There were also differences in the indirect effects involving FathEd and Profic. These results suggest that factors that affect test performance operate quite differently for male and female students and add to the authors' concerns over sex differences in test performance.

IMPLICATIONS

In this study the psychometric challenges of developing a college admissions test in Jordan for a highly selective sample of academically talented Jordanian students were examined. Although most items showed adequate variability and student performance did not depend on which form of the test they received, analyses using item response theory showed that several items did not exhibit adequate discrimination. Differential item functioning analyses also indicated that numerous items showed pronounced sex differences. Based on these results, approximately one-half of the test items need to be modified or omitted. However, it is likely that these findings are at least partly attributable to the selective nature of the sample.

The results of the path analysis showed that the initial path model was not supported for male students, and was only moderately supported by the results for female students. For males, there was no evidence that either attending a private school or coming from a family in which one or both parents had a formal education affected their test performance. For female students, the strongest predictors of student performance included parental educational level and whether students attended a government-sponsored school or a private school. In general, female students from families in which the mother had a formal education and who

attended a private school outperformed females from families with mothers with no formal education. This is not a surprising finding, but it is surprising that it does not extend to male students and their fathers.

Based on these findings, the authors have made the following recommendations to the expert panel for their consideration prior to the next piloting of the revised test: (1) A carefully constructed stratified random sampling scheme of at least 1,000 students should be adopted that provides appropriate representation from various subgroups, including sex and type of school attended. This is the simplest and most direct way of ameliorating most of the concerns over restricted variation in the data; (2) The 30 items flagged by the item analyses should be examined to determine whether they should be rewritten or deleted altogether and new items substituted. If new items are used it is important that their difficulty be appropriate for the students who will sit this examination. The panel should also consider whether the presence of a strong factor structure of the test is desirable (the authors think it is). (3) The strong evidence of differential item functioning for 19 items means that these items also require attention. Combined with the path analysis results for male and female students, it appears that eliminating sex differences in test performance is the next psychometric challenge in the test development process. (4) Variables should be better defined to allow the relationships among predictors and student proficiency to be more credibly examined. For the purposes of such examination, parental education should be defined as something other than a dichotomous variable.

The authors consider that adoption of these recommendations can assist the expert panel in producing an admissions test with strong psychometric properties that can play an important role in developing a skilled and well-educated labor Jordanian work force in the 21st century.

REFERENCES

- Abu-Lebdeh, K., & Ahlawat, A. K. (1997). *A comparative study of family background and Grade 4 students' achievement in math, science, and Arabic language before and after the implementation of education reform*. National Center for Human Resources Development, **53**, Amman, Jordan.
- Alnabhan, M. (1997). *Achievement level of Jordanian basic education students in science*. National Center for Human Resources Development, **50**, Amman, Jordan.
- Arbuckle, J. L. (1997). *Amos user's guide version 3.6*. Chicago: Small Waters Corporation.
- Embretson, S. E. (1996). Item response theory models and spurious interaction effects in factorial ANOVA designs. *Applied Psychological Measurement*, **20**(3), 201-212.
- Fischer, G. H. (1995). Some neglected problems in IRT. *Psychometrika*, **60**(4), 459-487.
- Glass, G., & Hopkins, K. (1996). *Statistical methods in education and psychology*. Mahwah, NJ: Prentice-Hall.
- Hambleton, R. K., & Swaminathan, H. (1985). *Item response theory: Principles and applications*. Boston: Kluwer-Nijhoff.
- Hatcher, L. (1996). *A step-by-step approach to using the SAS system for factorial analysis and structural equation modeling*. Cary, NC: SAS Institute.

- Hayduk, L. A. (1987). *Structural equation modeling with LISREL: Essentials and advances*. Baltimore, MD: Johns Hopkins University Press.
- Hu, L., & Bentler, P. M. (1995). Evaluating model fit. In R. Hoyle (Ed.), *Structural equation modeling: Concepts, issues, and applications*. Thousand Oaks, CA: Sage.
- Joreskog, K. G., & Sorbom, D. (1993). *Lisrel 8 user's reference guide*. Chicago, IL: Scientific Software.
- Kacmar, K. M., & Valle, M. (1997). Dimensionality of the measure of ingratiation behaviors in organizational settings. *Educational and Psychological Methods*, *57*, 314-328.
- Keyes, T. K., & Levy, M. S. (1997). Analysis of Levene's Test under design imbalance. *Journal of Educational and Behavioral Statistics*, *22*(2), 227-236.
- Mislevy, R. J., & Bock, R. D. (1997). *Item analysis and test scoring with binary logistic models*. Mooresville, IN: Scientific Software.
- Mueller, R. O. (1996). *Basic principles of structural equation modeling: An introduction to LISREL and EOS*. New York: Springer-Verlag.
- Muthen, B., Kao, O., & Burstein, L. (1991). Instructionally sensitive psychometrics: Application of a new IRT-based detection technique to mathematics achievement test items. *Journal of Educational Measurement*, *28*(1), 1-22.
- Schafer, L. L. (1997). *Analysis of incomplete multivariate data*. London: Chapman & Hall.
- SPSS Inc. (1997). *SPSS for Windows*. Chicago, IL: Author.
- Swaminathan, H., & Rogers, H. J. (1990). Detecting differential item functioning using logistic regression procedures. *Journal of Educational Measurement*, *27*, 361-370.