GMAC®

Graduate Management Admission Council®
Creating Access to Graduate Business Education®

Assessing Validity by Combining Examinees across Programs

Eileen Talento-Miller, Lawrence M. Rudner, & Kara Owens

GMAC® Research Reports • RR-06-14 • August 18, 2006

Abstract

Validity studies are a crucial part of any assessment program. For admissions tests, predictive validity evidence is commonly collected at the program level and aggregated by computing the mean or median validity coefficients across all programs. This approach documents overall test validity at the program level. However, in order to assess validity for groups, such as those formed by language, citizenship, or culture, investigations of validity at the examinee level would be more appropriate. However, the existence of program effects can make this approach problematic. This paper examines and compares estimating validity by meta-analyzing study-level results and compares the approach with combining data and assessing validity at the individual level. Alternate approaches are examined to account for program effects such as standardizing outcome measures, dummy coding for programs, and hierarchical linear modeling. The results of this study can help identify preferred methodologies for investigating cultural differences using large data sets aggregated across institutions.

Validity studies are conducted to demonstrate a relationship between performance on an assessment and the specified outcomes or purposes of that particular evaluation method. Ideally, a validity study would be conducted in every situation and for every population for which the test scores might be used. In practice, several situations or possible populations cannot be studied thoroughly because of limitations with sample size. In these cases, it is assumed that validity will generalize to similar situations based on meta-analytic results from previous studies. However, if a population, or subgroup of a population, has not been evaluated in previous studies, then no validity evidence exists for that group. As a result, any potential bias that might result from using the test to draw inferences about that particular population has not been determined. For instance, the validity of a given test administered in English may differ if the examinee's native language is not English. Is it sufficient to compare validity for the native and non-native English groups, or might there be additional differences depending on the specific native tongue of the examinee? A number of non-U.S. schools use admission tests developed in the United States and administered in English. The Graduate Management Admission Test® (GMAT®) exam is an example. Although developed in the United States, this exam is used for programs around the world, and approximately 45% of examinees are not U.S. citizens (GMAC®, 2005). Because of the number of native languages that might be present, it would be difficult to conduct a separate validity study for each language represented in any one program. However, if data from that program could be combined with data from other programs, there may be enough cases to evaluate potential differential validity and differential prediction by native language groups.

In the case of admissions testing, hundreds of validity studies are conducted to ensure that the test scores are related to later performance in the program to which the individual has applied. Oftentimes, validity coefficients and regression lines are used to evaluate the validity of admissions test scores for selecting students that will be successful in a graduate or undergraduate program (Young, 2001). Because of the wealth of data available on different programs, generalizability to different situations can be established. However, less information is available when generalizing validity results to different student populations. As with the previous example, large groups can be analyzed and examined for differential validity and differential prediction; however, smaller groups are

generally combined even though there may be more variability within the group than between groups.

There have been several methods used to combine study data or individual data to summarize validity for a test. Meta-analyses generally use mean or weighted-mean study results, as demonstrated in Kuncel, Crede, and Thomas (2004), or median study values, such as those reported in Talento-Miller and Rudner (2005). Because validity is program specific, average program-level (PL) results would be the best indicator of the expected validity of a test. The ideal measure for a group, such as a specific language or citizenship group, would be the mean or median PL validity computed just on the data of the target group in each program. As previously stated, this is impractical for many groups in most programs.

Some studies combine data at the individual level to assess validity for smaller groups, but recognizing differences at the program level, attempt to compensate using different methods. Many methods include adjustments to grades as the success criteria to obtain more accurate prediction and allow for comparisons across programs (Young, 1993). For instance, Braun and Szatrowksi (1984) developed a strategy for combining universal criteria from similar institutions to adjust grades and more accurately reflect performance. The scale-linkage algorithm uses data collected from students who applied and were accepted to the same two schools but attended only one of the schools. Data on performance in the attended program and admission criteria, such as undergraduate grades, were used to adjust for grading differences at the two institutions and allow for data to be combined universally and validity to be estimated across institutions.

Sireci and Talento-Miller (2006) also combined data across programs to analyze gender groups and different racial/ethnic categories. Recognizing that differences may exist in grading standards across the various programs, the outcome variable—first-year grades—was standardized within school before conducting the analyses. Although this method helps to equalize variance in grades across programs, it does not control for other systematic differences among programs. Talento-Miller (2006) went one step further and not only standardized grades within school but also added program effect in the model by dummy coding the programs and including the variables in the prediction equations. Because only six programs were studied, the addition of the five dummy variables still allowed for robust analyses with the sample size available.

A recent study by Brown and Zwick (2006) accounted for the multilevel nature of the data by using hierarchical linear modeling to analyze the validity of admission factors across several schools. This more complicated methodology, however, does not provide an easy method for estimating the magnitude of the validity coefficient, and models were instead compared by differences in variance components.

The purpose of this study was to compare methods summarizing PL validity results to methods examining validity assessed at the individual level (IL). If individual data can be used without introducing systematic unexplained variance, then studies can be conducted analyzing validity for small groups by combining data across multiple situations.

Methods

Data

The study was conducted by combining data from 163 individual validity studies conducted for graduate management programs between 1997 and 2004, with a total of 20,270 cases. Each of the studies was conducted for the schools by the Graduate Management Admission Council®, which offers a free Validity Study Service (VSS) to any program that uses the GMAT® exam as part of its admission process. Because all the studies were conducted using the same variables and the same methodology, the error due to study differences is minimized. This provides an opportunity to assess the generalizability of the validity of the predictors and estimate the amount of variance attributable to program differences.

Variables

The VSS asked schools to submit GMAT® scores and undergraduate grade point average (GPA) to predict the outcome variable of first-year or mid-program GPA. Although every case included the outcome variable, some cases were missing various predictors. In particular, undergraduate GPA was not always available, especially for non-U.S. students, and some schools did not collect information on the GMAT® Analytical Writing Assessment (AWA) scores. Study outcomes that were

analyzed for this research include the results of 12 analyses representing the simple and multiple correlations of predictor combinations. The variables used in the 12 analyses are listed in Table 1. GMAT® Total scores include performance on both the verbal and quantitative sections, but not the writing section. As a result, predictor combinations that include total scores would not also include verbal and quantitative scores, but could include AWA scores.

Table 1. Predictor Combinations Used for VSS Studies

Abbreviations	Variables
AWA	GMAT® Analytical Writing Assessment
UGPA	Undergraduate Grade Point Average
GMATV	GMAT® Verbal Scores
GMATQ	GMAT® Quantitative Scores
GMATT	GMAT® Total Scores
VQ	Verbal + Quantitative
VQA	Verbal + Quantitative + Analytical Writing
TA	Total + Analytical Writing
VQU	Verbal + Quantitative + Undergraduate GPA
TU	Total + Undergraduate GPA
VQAU	Verbal + Quantitative + Analytical Writing + Undergraduate GPA
TAU	Total + Analytical Writing + Undergraduate GPA

Data Analyses

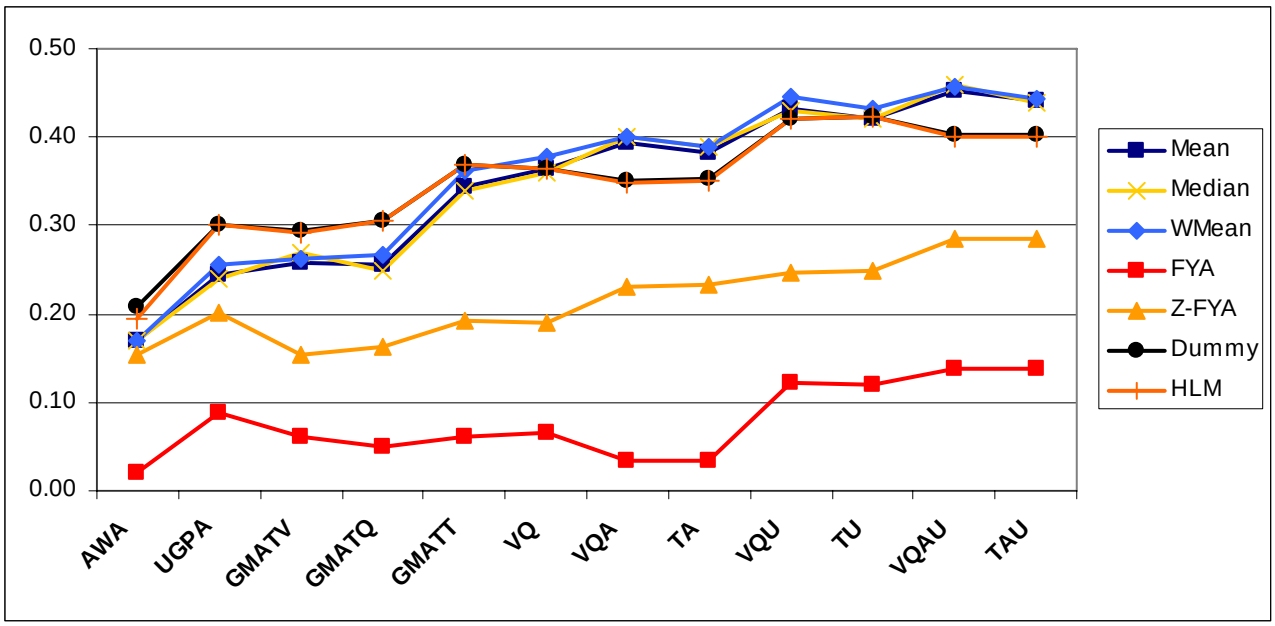
The analyses were conducted to determine whether differences existed in average validity using various methods. The PL results for each study were summarized using means, medians, and means weighted by sample size. Four methods were used to calculate IL validity. First, results for each of the predictor combinations were calculated by using ordinary least squares (OLS) regression to predict first year average (FYA). For the remaining three analyses, the dependent variable of FYA was standardized within the study. The second analysis used OLS regression to predict standardized first year average (Z-FYA). For the third analysis, a study effect was entered into the OLS regression by including dummy codes representing the study or program effects in the prediction equations for each predictor combination. Hierarchical linear modeling (HLM) was selected as the final analysis approach. For this procedure, predictors were entered as level-one variables and study was entered as the level-two variable. To compare the results of the HLM analyses to the OLS regression analyses, an estimate of R was calculated as the square root of the proportion of variance explained based on distance measures explained in Roberts and Monaco (2006).

Bootstrapping methods were used to further compare PL versus IL summary method results to examine the effect of the number of studies being compared. Each sample was created by randomly drawing studies from the original 163 studies without replacement. Fifty independent sets of 5, 10, 15, 25, and 50 studies were drawn and analyzed. For each sample, PL weighted-mean validity and IL validity, calculated using OLS with dummy codes, was compared for predictor combinations. Because a number of schools did not include AWA information in their studies, it was difficult to obtain samples that contained complete data for this variable; therefore combinations with AWA were excluded from this analysis.

Results

The results are presented graphically in Figure 1. Results for each of the PL and IL methods are plotted for each predictor or set of predictors listed in Table 1. It is apparent from the close proximity and overlap of the lines for the mean, median, and weighted-mean lines on the graph that the three PL summary statistics produced very similar results. Results at the IL using FYA show that differences among programs in grading can severely inhibit the measurement of predictive validity, since the values observed for this method are much lower than any of the other IL or PL methods. Once grades are standardized within program, Z-FYA, the validity values at the IL are closer to the PL, yet still noticeably repressed. Both the OLS with dummy coding and the HLM-IL analyses appeared to be reasonable approximations of the PL results, with little distance between these two lines and the PL summary lines. In fact, the two analyses yielded estimated R values that were nearly identical with differences at most of 0.002, with only one exception. (The difference in R for the prediction using only AWA was 0.014.) In essence, dummy coding each of the studies allows there to be a unique intercept for the set of cases included in any one study, which is similar conceptually to the HLM analyses. Although these IL analyses yielded results similar to the PL data, there appeared to be a pattern where the IL results were higher than PL for the individual predictors, but lower for the combinations of predictors.

Figure 1. Comparison of Methods for Calculating Validity of Predictors and Combinations



The comparison of methods for summarizing the validity results at the PL and at the IL was based on data from 163 studies. To explore whether similar findings would be observed if fewer studies were involved, a series of random draws of varying size from the database of studies was used to replicate the results. Weighted-mean validity PL results were compared to the IL procedure using OLS regression on Z-FYA with dummy coding.

The results are summarized in Table 2. The entries represent the mean of the validity values observed across multiple draws at each of the sample sizes. For instance, across more than 40[1] draws of five studies, the average of the weighted-mean validity for UGPA is 0.167. The same draws of five studies for the dummy-coded validity

[1] Because some cases had missing UGPA data, there were less than 50 draws for all combinations including UGPA.

method yielded an average of 0.245. Lower validity values were observed across all results compared to the aggregate results for all 163 studies due to the imposed limits on variability by drawing only subsets of the original dataset. In the table, the method yielding the larger validity value is represented using bold type. The final column represents the difference between the PL and IL methods, with negative values indicating that the IL method resulted in higher validity estimates. Examining the final column, it is apparent that the largest differences between the methods were observed for the single predictors (UGPA, Verbal, Quant, and Total), where the IL validity coefficients consistently exceeded the PL values for validity. When predictors were combined, the two methods appeared much more compatible with smaller differences on average. Additionally, for the VQU combination there was no clear pattern of one method producing higher results. Although it appears that the number of predictors in the model affects the similarity of the results, this hypothesis would need to be tested systematically with a greater number of variables[2].

Table 2. Comparison of Mean (SD) Validity from Bootstrapping at Program-Level and Individual-Level

Variables		5	10	15	25	50	*Total*	PL-IL
UGPA	PL	0.167 (0.07)	0.168 (0.04)	0.176 (0.02)	0.177 (0.02)	0.172 (0.01)	*0.171 (0.03)*	-0.081
	IL	**0.245 (0.07)**	**0.244 (0.05)**	**0.260 (0.03)**	**0.254 (0.02)**	**0.260 (0.02)**	***0.252 (0.04)***	
Verbal	PL	0.210 (0.05)	0.204 (0.04)	0.207 (0.03)	0.205 (0.02)	0.205 (0.01)	*0.206 (0.03)*	-0.068
	IL	**0.273 (0.06)**	**0.268 (0.05)**	**0.279 (0.05)**	**0.274 (0.03)**	**0.278 (0.02)**	***0.274 (0.04)***	
Quant	PL	0.185 (0.06)	0.185 (0.03)	0.173 (0.03)	0.179 (0.02)	0.175 (0.01)	*0.179 (0.03)*	-0.070
	IL	**0.248 (0.07)**	**0.246 (0.05)**	**0.248 (0.04)**	**0.250 (0.03)**	**0.251 (0.02)**	***0.249 (0.04)***	
Total	PL	0.253 (0.05)	0.254 (0.03)	0.248 (0.04)	0.251 (0.02)	0.247 (0.01)	*0.251 (0.03)*	-0.052
	IL	**0.305 (0.06)**	**0.301 (0.04)**	**0.304 (0.05)**	**0.303 (0.03)**	**0.304 (0.02)**	***0.303 (0.04)***	
VQ	PL	0.293 (0.05)	0.288 (0.03)	0.281 (0.04)	0.284 (0.02)	0.281 (0.02)	*0.285 (0.03)*	-0.021
	IL	**0.311 (0.06)**	**0.304 (0.04)**	**0.304 (0.05)**	**0.304 (0.03)**	**0.305 (0.02)**	***0.306 (0.04)***	
VQU	PL	0.344 (0.06)	**0.337 (0.04)**	0.335 (0.03)	**0.341 (0.02)**	0.336 (0.01)	***0.339 (0.04)***	0.001
	IL	0.344 (0.06)	0.335 (0.04)	**0.336 (0.04)**	0.338 (0.02)	**0.337 (0.02)**	*0.338 (0.04)*	
TU	PL	0.311 (0.06)	0.311 (0.04)	0.312 (0.04)	0.316 (0.02)	0.310 (0.01)	*0.312 (0.04)*	-0.024
	IL	**0.338 (0.06)**	**0.333 (0.04)**	**0.337 (0.04)**	**0.337 (0.02)**	**0.337 (0.02)**	***0.336 (0.04)***	

Discussion

The comparison of methods for summarizing the results of validity studies showed that across a large number of studies, similar values were observed for several methods. Assuming the results from the studies form a normal distribution, comparing the PL methods (mean, median, and weighted-mean) would be expected to give similar findings. For IL methods, it is clear that program effects must be considered in order to observe values such as those that would be encountered for each program. It is not sufficient to standardize the outcome variable, but additional program-level variance can be explained by either including the programs as dummy variables in an

[2] Combinations using AWA lend support to this hypothesis, but because few draws were possible with complete data, the results are not presented.

OLS regression or by modeling the PL effects in an HLM analysis.

The weighted mean, which weighted studies at the level of the individual, and the dummy-coded programs were further compared in a number of replications. The fact that the IL results yielded consistently higher validity estimates for individual predictors, and in many cases for the multiple predictions, is likely a result of the increased variance in the predictors. Within any one school, restriction of range due to the admission selection process limits the variability of the predictors and, therefore, makes it more difficult to explain differences in the criterion using those predictors. When the data from the schools are combined, a greater range of scores on the predictors are represented, and the greater variability makes it possible to explain more differences. The effect appears to be reduced when more variables are introduced into the model, since the intercorrelations among the variables explain some of the additional variance.

The implications of the study suggest that it is possible to obtain reasonable validity estimates when combining data across multiple situations if accommodations are made to ensure that the additional program-level variance is accounted for, such as by including dummy variables in the model or modeling the levels in the data hierarchically. From a practical standpoint, it is more difficult to analyze data using dummy variables when a large number of studies are involved. For instance, the use of dummy coding in the present study meant the addition of 162 predictors for each of the analyses. HLM analyses are more robust with the greater number of studies, but interpretation may be more difficult since a measure of explained variance is not as readily available. For these reasons, dummy coding would be more practical if combining data for a small number of studies, while HLM may be more useful for large numbers of studies.

Although the IL values may be considered reasonable approximations of the expected validity for groups, there appears to be systematic differences between IL and PL findings. One of the benefits to combining data and analyzing results at the IL is the ability to compare groups with more reasonable sample sizes than may be available within a single study. Groups can be compared using the same method (male versus female using HLM), but it would not be appropriate to compare the results across PL and IL methods (PL mean meta-analytic findings for male students versus IL results for female students). Future research can further examine the differences between dummy coding and HLM analyses. Another area that can be further explored would be the over-estimates of PL validity using the IL results for single and multiple predictors. If dummy coding of programs adds to the explained variance, then could other dummy codes be added to identify additional differences among cases? For instance, what would be the effect of including additional dummy codes for particular concentrations, such as finance or marketing? This is another direction for future research with the effect of such a change to the model compared to average program results.

There are several limitations to the present study. The study examined different methods for calculating validity to describe how combined results can contribute to the generalizability of validity in different situations or for different groups. One important distinction between many of the validity generalizability studies and the current study is the correction for known statistical artifacts (Murphy, 2003). For instance, meta-analyses such as the one by Kuncel et al. (2004) may correct study results for restriction of range, sampling error, and criterion reliability. None of these corrections were utilized with the values in the current study. Therefore, the validity values observed here for the PL results would be expected to be higher if corrected to reflect actual expected validity for the entire applicant population rather than the admitted students. Methods would need to be developed or adapted to correct the results of IL methods to identify the best estimates for the true validity values. Another limitation of the study lies in the nature of the database. The database of studies had the unique property that all the studies were conducted in the same way, but because the programs elected to participate in the VSS, they cannot be assumed to be a random sample. There were limitations imposed on participating schools (such as a minimum sample size) that affect representation. Furthermore, a program may be represented multiple times in the sample if they conducted studies in different years. The current study examined seven methods, but there may be other reasonable PL or IL methods that were not examined here, such as empirical Bayes analyses.

The comparison of methods for examining validity across different programs and different studies showed that it is possible to combine data for more robust analyses without increasing the error variance by unreasonable amounts. It is hoped that this research can be replicated in additional samples and extended by looking at additional methods. Importantly, ways to approximate true validity by accounting for possible statistical artifacts affecting PL or IL validity estimates need to be carefully considered to have the most informative view of the implications of assessments.

Contact Information

For questions or comments regarding study findings, methodology or data, please contact the GMAC Research and Development department at research@gmac.com.

Acknowledgements

A version of this paper was presented at the International Test Commission's 5th International Conference on Psychological and Educational Test Adaptation across Languages and Cultures, July 6-8, 2006, Brussels, Belgium.

References

Braun, H. I., & Szatrowski, T. H. (1984). The scale-linkage algorithm: Construction of a universal criterion scale for families of institutions. *Journal of Educational Statistics, 9*, 311-330.

Brown, T., & Zwick, R. (2006). Application of hierarchical linear modeling to a predictive validity study of college admission tests. Paper presented at the annual meeting of the National Council on Measurement in Education, April 8-10, 2006, San Francisco, CA.

Graduate Management Admission Council®. (2005). *Profile of GMAT® Candidates, 2000-01 to 2004-05.* McLean, VA: Graduate Management Admission Council®.

Kuncel, N., Crede, M., & Thomas, L. (2004). A comprehensive meta-analysis of the predictive validity of the Graduate Management Admission Test (GMAT) and undergraduate grade point average (UGPA). Presented at the Annual Meeting of the Society for Industrial and Organizational Psychology, April 2-4, Chicago, IL.

Murphy, K. (Ed.). (2003). *Validity generalization: A critical review.* Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

Roberts, J., & Monaco, J. (2006). Effect size measures for the two-level linear multilevel model. Paper presented at the annual meeting of the American Educational Research Association, April 7-11, 2006, San Francisco, CA.

Sireci, S., & Talento-Miller, E. (2006). Evaluating the predictive validity of Graduate Management Admission Test scores. *Educational and Psychological Measurement, 66*, 305-317.

Talento-Miller, E. (2006). GMAT® validity for six European MBA programs. GMAC® Research Report RR-05-07. McLean, VA: Graduate Management Admission Council®.

Talento-Miller, E., & Rudner, L. (2005). GMAT® validity study summary report for 1997 to 2004. GMAC® Research Report RR-05-06. McLean, VA: Graduate Management Admission Council®.

Young, J. L. (1993). Grade adjustment methods. *Review of Educational Research, 63*, 151-165.

Young, J. L. (2001). Differential validity, differential prediction, and college admission testing: A comprehensive review and analysis. The College Board Research Report No. 2001-6. New York, NY: The College Entrance Examination Board.