

Assessing the internal structure of math exam scores with confirmatory factor analysis

Comprehensive Exam Project

Maggie Farley

University of New Mexico

Committee Members

Chair of Committee: Dr. Gary Ockey

Dr. William Bramble

Dr. Scott Marley

Dr. Jay Parkes

Introduction

In the current age of educational reform, the degree to which test scores measure what is being taught in the classroom carries significant consequences for schools facing accountability measures (La Marca, 2001). The movement in the USA to establish a level playing field for all students has led some states to develop equitable assessments for students not fully proficient in English. Providing a test in the student's native language is an accommodation that is routinely used by a number of states (Stansfield, 2003). The Public Education Department (PED) of New Mexico is conducting studies that address reliability and validity concerns with state assessments given in English and in Spanish.

The Standards for Educational and Psychological Testing call for construct validity evidence of the internal structure of a test (AERA, APA & NCME, 1999). The degree to which we are able to validly interpret the test scores is dependent on our understanding of the internal structure of the exam. La Marca (2001) contends that the validity of test score interpretation depends on establishing methodological requirements for the alignment of exams with the content of the subject areas being assessed. The breadth of content covered in state assessments might limit the measurement of the multiple content standards and objectives being assessed (La Marca).

Confirmatory factor analysis (CFA) provides a framework with which to measure how well test items are aligned with the benchmarks or standards being tested. One CFA study by Stevens, Farley, & Zvoch (2006) examined TerraNova 2001 scores for a sample of 6,692 New Mexico 8th grade students. The TerraNova exam is an achievement test that evaluates students in the following four content areas: mathematics, language arts,

reading, and science and social studies. The TerraNova exam measured the proposed four constructs with a total of nine subtests. This study compared factor structures based upon exploratory factor analysis, previous research, and CFA modification indices, to the four-factor structure proposed by the TerraNova exam. The results showed that the proposed model based upon the four constructs had the best overall fit. However, a more parsimonious three-factor model that combined the measures of reading with the measures of social studies and science showed comparable fit.

An additional purpose of the study was to examine the internal structure of the exam with respect to the discriminant validity of the scores on the proposed four factors of knowledge. Stevens et al. (2006) found high intercorrelations ranging from .83 to .98 between the four factors on the TerraNova and concluded that a valid interpretation of the test scores would not regard the constructs as representing different facets of achievement. Based upon the results of this study, Stevens et al. concluded that further empirical research of the internal structure of high stake state assessments was needed in order to ensure the correct interpretation of test scores.

Construct Validity

Discriminant and convergent validity are two types of construct validity commonly evaluated by CFA. Evidence for discriminant validity includes correlations between factors that are not high when the factors should measure distinct constructs. If high intercorrelations are not found between the factors, then there is evidence of discriminant validity. In contrast, relatively high factor loadings on variables the factor is purported to measure are evidence of convergent validity. If the intercorrelations between indicators that load on the same construct are at least moderate in magnitude,

there is evidence of convergent validity (Kline, 2005). In this study, CFA was used to model the internal structures of the 2006 New Mexico 8th grade math state assessments in English and Spanish. The purpose of this study was to determine the extent to which there is evidence in these exams of the constructs that the tests claim to measure.

Structural Equation Modeling

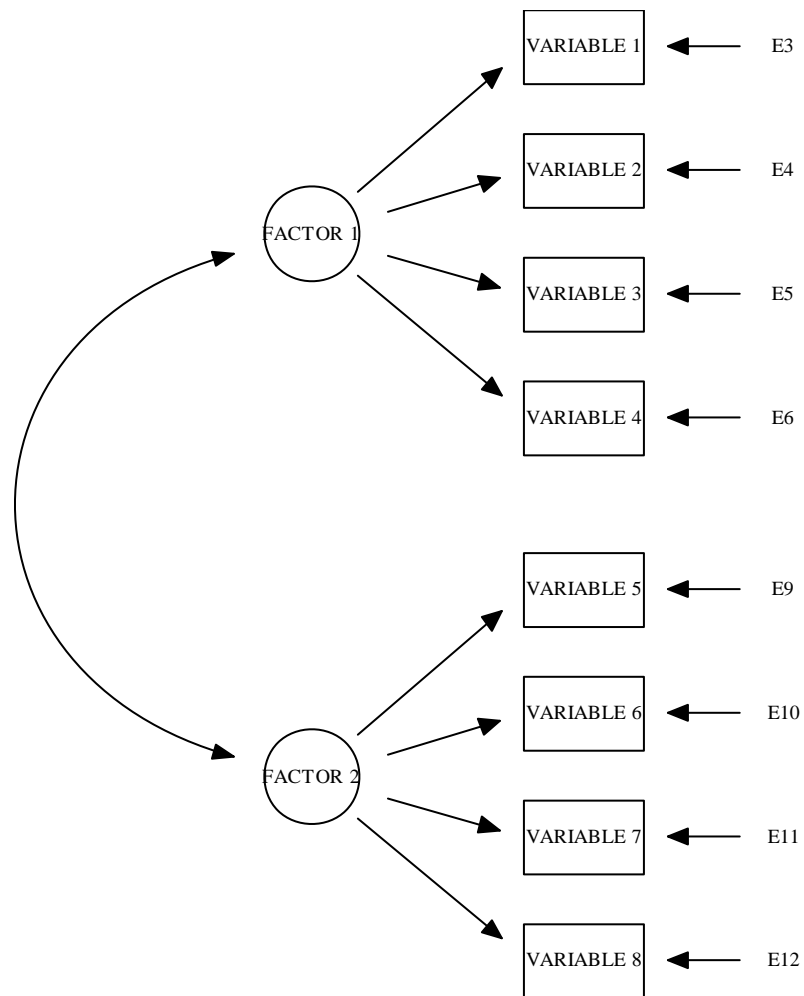
The variable of interest in social science research is often hypothesized to be known through its relationship to variables that can be measured. In the structural equation modeling (SEM) framework, the hypothesized constructs are also called *factors*, *unobserved* variables or *latent* variables, while the measured variables are also referred to as *observed* variables, *indicators* or *manifest* variables. Unlike other statistical methods, SEM allows for the inference of relationship at the construct level (Byrne, 1994). It is assumed that the correlation between the measured variables and the construct is being driven by what the measured variables have in common with regard to the latent variable; thus, the construct is viewed as predicting the measured variables. In CFA, a series of measurement equations for the model make the causal relationship between variables explicit.

SEM allows the researcher to diagram the complex nature of hypothesized relationships between constructs and measured variables. In SEM, error in measurement is distinguished from the disturbance, or residual in the prediction of a factor. An important characteristic of SEM is that it incorporates both measured and unmeasured variables in the measurement of error whereas other multivariate procedures do not have the capacity of estimating measurement error (Byrne, 1994).

EQS (Bentler, 1985-2006) is based on the Bentler-Weeks representation system in

which all variables are classified as either dependent or independent. Confirmatory factor analysis is strictly based on measurement models and hence causality is not inferred between latent factors. Thus, in CFA the latent variables are independent variables and the observed variables are dependent variables. Another important component of the Bentler-Weeks model is that the covariances between observed variables are not estimated but remain to be explained by the model (Byrne, 1994). In contrast, the variances of the factors and the errors are parameters that are estimated in ascertaining degrees of freedom for the model (Byrne).

Figure 1. Sample confirmatory factor analysis model diagram



Methodology

A path diagram of a model analyzed in this study is shown in Figure 1 for the purpose of illustration. The unidirectional arrows are aiming at the dependent, and in this study, measured variables. The one-way arrows pointing from the errors indicate the impact of random measurement error and the variance not accounted for by the factor loadings on the observed variables. Curved two-way arrows represent correlations between pairs of variables. Observed, or measured variables, are enclosed in rectangles while unobserved variables are enclosed by ellipses or circles.

Overview

A dual language school in the Albuquerque Public School district administered the standards-based assessment to the same students in English and Spanish in the spring of 2006. The exam was given to students in grades 3 through 9 in the subject areas of math, reading, science and writing. The analysis of the Spanish and English exam scores is being utilized by the PED to inform a future statewide validity study.

The cleaning and screening of the test score data received from the PED involved the selection of students who had completed the reading and math exams in both English and Spanish. Alpha numeric symbols representing scores were converted to the numerical equivalent of the score as indicated in a test scoring sheet provided by the district administering the exam. Correlations between the math and reading exams taken in English and Spanish for grades 3 through 9 were evaluated. However, reading scores were not considered for analysis in this study due to a lack of identical items on the English and Spanish versions of the reading exams. The sample size of students who had taken the math exam in both English and Spanish, as well as, correlations on the raw total

scores on the math exam were criteria used to select 8th grade as a sample population. The analysis of math scores at all grade levels went beyond the scope of this study.

Participants

A screening of the 8th grade math scores revealed that a population of one-hundred-and-twenty-three students completed the exam. However, only sixty-seven students completed the math exam in both English and Spanish. All sixty-seven students in the sample population identified themselves as Hispanic. Thirty-six students were female and thirty-one students were male. Twenty-three students were currently in an English Language Learner (ELL) program, whereas eleven students had been exited for one to two years and thirty-three students had been exited for four years. Sixty-five students were identified as non-immigrant and two students were identified as having immigrant status. The 8th grade math scores in English and Spanish for the sample of sixty-seven students were the basis for the factor analyses conducted in this study.

Structure of 8th Grade Math Exam

The math exam had an internal structure that tested students on five standards. These five standards were: 1) Numbers & Operations, 2) Algebra, 3) Geometry, 4) Measurement and 5) Data Analysis & Probability. In turn, each of the five standards was composed of benchmarks that targeted specific areas of learning. A list of the five standards and their benchmarks are shown in Table 1. The Algebra and Data Analysis & Probability benchmarks analyzed in this study are in bold.

Table 1. 8th Grade Math Standards and Benchmarks

Name of Math Standard

Numbers & Operations -- Benchmarks

Number Sense

Operations

Computation and Estimation

Algebra -- Benchmarks

Patterns & Functions

Algebraic Symbols

Modeling Relationships

Analyzing Changes

Geometry -- Benchmarks

Analyzing Shapes

Coordinate Geometry

Transformations and Symmetry

Spatial Sense and Modeling

Measurement -- Benchmarks

Measurement Systems

Applied Measurement

Data Analysis & Probability -- Benchmarks

Data Collection and Representation

Statistical Methods

Data Analysis

Probability

Review of Data and Data Analyses

The types of test items included multiple-choice, short-answer and open-ended. The corresponding scoring scales were (0-1) for multiple-choice items, (0-2) for short-answer items and (0-4) for open-ended items. A closer look at the item scores for each benchmark revealed that of the five standards, only two standards had item scores for each of the listed benchmarks. A five-factor model was not constructed due to the fact that three out of the five standards listed above did not have a measurable benchmark structure. In CFA, at least two indicators are needed for each factor in order to specify or fit the model. Thus, an assessment of validity in math exam scores exclusively at the standard level was not possible.

Although a full five-factor structure for the 8th grade math exam was not indicated, a two-factor structure built on the aggregation of item scores at the benchmark level for the standards of Algebra and Data Analysis & Probability was analyzed for discriminant validity. As seen in Table 1, the four Algebra benchmarks treated as variables in the factor analysis were, A) Patterns & Functions, B) Algebraic Symbols, C) Modeling Relationships, and D) Analyzing Changes. The four Data Analysis & Probability Benchmarks analyzed in the factor analysis as variables were, A) Data Collection & Representation, B) Statistical Methods, C) Data Analysis and D) Probability.

A second purpose of this study was to evaluate the correlation of item scores on a standard's benchmarks for evidence of construct validity. An evaluation of the item level scores revealed that the total number of items and points differed on the Spanish and English versions of the math exams in all but two standards, Algebra and Data Analysis

& Probability. Algebra was given preference over Data Analysis and Probability because it was felt that the Algebra scores would be more representative of math ability at the 8th grade level for the majority of students. In addition, the Algebra benchmarks had a high number of common items on each of its four benchmarks.

The item sequences for the English and Spanish exams were evaluated and items not in common were removed from the analysis of the Algebra benchmarks. Three item scores (numbers 4, 7 & 9) were removed from the evaluation of the Spanish Algebra 8th grade test scores, and two item scores (numbers 9 & 10) were removed from the evaluation of the English Algebra 8th grade test scores. The four Algebra benchmarks consisted of a total of seventeen items; the item scores reflected nine multiple-choice items, seven short-answer items and one open-ended item.

Missing Data

The missing data in this study was assumed to be missing at random (MAR), which meant that it was assumed that incomplete cases differed from complete cases by chance (Kline, 2005). The percentages of missing data for each of the seventeen items are provided in Table A13 in Appendix A. Missing data was imputed by EQS using expectation-maximization (EM), which was composed of two iterative steps. First, regression of the missing variables on the remaining variables estimated missing observations for each case with predicted scores (Kline, 2005). In the maximization step, the covariance and mean matrices were estimated from the imputed data and the covariance matrix was corrected with respect to the imputed means using maximum likelihood (ML) estimation (Bentler, 2006). The two steps were repeated until the overall parameter estimates converged (Bentler). After EM imputation was completed, data

values were manually rounded to the nearest whole number in view of the categorical nature of the data.

Assumptions

The method of maximum likelihood estimation in standard confirmatory factor analysis assumed multivariate normal distributions for the indicators (Kline, 2005). An examination of scatter plots (see Figures B1-B16 in Appendix B) revealed positive linear relationships between the benchmark variables and the factors. A second property of multivariate normality was that the univariate and joint distributions for the measured variables were normally distributed. Although it is often impractical to check the bivariate distributions of all variables, the inspection of univariate distributions can pinpoint many issues with multivariate normality (Kline).

Skew and kurtosis are two ways in which a distribution can be nonnormal, however, the measure of kurtosis is much more relevant to the assumption of multivariate normality in structural equation modeling (Bentler, 2006). Skew is a measure of the symmetrical distribution of the data above or below the mean, whereas kurtosis is an index that measures the normality of the data with regards to the narrowness or width of the peak and the tails of the distribution. Under normality, skew measures should be close to 0.00, whereas kurtosis indices outside the range of -3.00 to 3.00 might indicate negative or positive kurtosis (Kline).

Using SPSS 13.0 (1989-2006), skewness and kurtosis measures, as well as, histograms were examined for evidence of univariate normality (see Tables B1 and B2 in Appendix B). The univariate measures of skew and kurtosis seemed well within normal range. In addition, a view of the histograms did not indicate any obvious problems with

univariate skew or kurtosis (see Figures B17-B32 in Appendix B).

Mardia's multivariate kurtosis coefficients were reported for the models in Series 1. Mardia's coefficient reports the extent to which the tails of all measured variables (with the same component means, variances and covariances) differ from ones characteristic of a normal distribution. If the distribution only deviates marginally from that of a normal distribution, Mardia's coefficient will be close to 0.00 with a non-significant normalized estimate. Mardia's values outside the range of -3.00 to $+3.00$ indicate a departure from multivariate normality (Bentler, 2006). The possibility that outlying cases were contributing to the estimate of multivariate kurtosis was examined for all models. The measures of multivariate normality did not give any reason to question the use of normal theory statistics for the models in Series 1.

Model Series

EQS 6.1 for Windows, (Bentler, 1985-2006) is a structural equation computer program that was utilized in this study to evaluate the validity of the internal structure of the 8th grade math exam. The first series of CFA models examined the 8th grade math exam scores at the benchmark level with the item scores aggregated. The second series of CFA models examined the correlation of item scores with a selection of the four Algebra benchmarks. Analyses for the Spanish and English versions of the math exam scores were conducted separately for the sake of clarification in the evaluation of internal structural validity.

Series 1: Algebra and Data Analysis & Probability Benchmark Models

The English and Spanish models in Series 1 were designed to assess the internal structural validity of the benchmark structure for two standards, Algebra and Data

Analysis & Probability. The models in Series 1 were constructed first as two-factor models in order to assess discriminant validity. The Algebra standard was one of the factors with four loadings, which were represented by each of the four Algebra benchmarks. The second factor was the Data Analysis & Probability standard with four loadings or Data Analysis & Probability benchmarks. Correlations above .90 were found between the two factors in both of the English and Spanish models, and thus the two-factor models were collapsed into one-factor models representing General Math Knowledge. Evidence for convergent validity was then evaluated in the one-factor models of General Math Knowledge.

Series 2: Algebra Item Score Models

The English and Spanish models in Series 2 were designed to examine the correlations of item scores with each of the four Algebra benchmarks. Due to the small sample size, the models in Series 2 were run only as one-factor models. Evidence found in five out of the eight models indicated the likelihood of Heywood cases, which are inadmissible solutions in confirmatory factor analysis (Kline, 2005). Estimated correlations greater than 1.0 between the factor and the indicator pointed to the presence of one or more negative variance estimates in the model, which was an illogical problem making the converged solution inadmissible (Kline). Problems with Heywood cases may occur in studies that have small samples and extremely low population variances (Chen, Bollen, Paxton, Curran, & Kirby, 2001). The descriptive statistics of all items are provided in Appendix A for the sake of illustrating the problems encountered due to issues regarding sample size, low intercorrelations and categorical data analysis.

One-factor models will be evaluated for convergent validity for two English

benchmarks, Patterns & Functions and Algebraic Symbols, as well as, the Spanish benchmark, Patterns & Functions.

Model Identification

The aim in structural equation modeling is to specify a model that has a unique solution for the values of the parameters in the model so that the model can be tested (Byrne, 1994). Structural equation models may be just identified, overidentified, or under-identified. In an underidentified model the number of estimable parameters is greater than the number of observed covariances and variances of the observed variables, which allows for more than one possible solution for the parameters. Thus, in underidentified models the model cannot be evaluated empirically because attainment of consistent estimates for the parameters is not possible (Byrne). In contrast, when the number of data variances and covariances is equal to the number of estimable parameters, the model is considered to be just-identified. While the just-identified model is capable of yielding a unique solution for all parameters, it has no degrees of freedom and thus cannot be rejected (Byrne). Finally, an overidentified model is scientifically interesting because the number of estimable parameters is less than the number of data points. An overidentified model results in positive degrees of freedom, which means it can be rejected. In SEM, models are specified so that they meet the criterion of overidentification. Statistical indices of fit are then utilized by the researcher to determine whether the covariances of the proposed model are valid estimates of the observed data.

The specification of an overidentified model is one aspect of correctly specifying a SEM model. In addition, the parsimony of the model is an important consideration in

the evaluation of good model fit. Paths should not be added with the sole purpose of improving model fit. Another requirement of identification is that the scale of measurement be determined for the latent variable. The scale of measurement can be established by constraining the estimate of an indicator or factor loading mapped onto the latent variable to a known value such as 1.0.

Maximum Likelihood Estimation

Maximum likelihood estimation (ML) is a method of deriving parameter estimates that assumes a population distribution that is multivariate normal for the measured variables (Kline, 2005). The goal of ML is to minimize the residuals between the observed covariances and the covariances of the hypothesized model and thus maximize the likelihood that the estimated parameters are drawn from the population.

Robust Statistics and Goodness of Fit Indices

The chi-squared test yields a probability for the likelihood of the sample data obtained in the study if the model were correct in the population (Bentler, 2006). Thus, a significant chi-squared value indicates that if the model were correct, the sample would be unlikely to be observed. If the underlying assumption of normality is violated in maximum likelihood estimation, the normal theory chi-squared test statistic may not hold. The use of an estimation method that assumes an underlying normal distribution with a test statistic that has been corrected to take nonnormality into account has been found to be more appropriate than a mode of estimation that simply does not assume normality when evaluating model fit (Byrne, 1994). When distributional assumptions are violated, the Satorra-Bentler scaled statistic incorporates the model, estimation method, and sample kurtosis values into a scaling correction for the chi-squared statistic (Byrne).

The scaling correction is designed to adjust the mean of the sample distribution with a multiplier to yield a new statistic whose expected value, or mean, is closer to that of the chi-square distribution (Bentler).

Results should take more than one fit index into account because different fit indices describe different aspects of the model fitting process (Kline, 2005). Two widely used fit indices are the comparative or incremental fit index and an absolute misfit index, the root mean square error of approximation (RMSEA). The comparative fit index (CFI) is derived from a comparison of fit of the hypothesized model with a null model, which assumes that all correlations among variables are zero. Values for the CFI range from zero to 1.00. A fit value greater than .95 indicates very good fit and fit values between .90 and .95 are considered acceptable (Hu & Bentler, 1999). The RMSEA is a measure of the lack of fit of the researcher's model to the population covariance (Kline, 2005). In other words, RMSEA compares the lack of fit in a model to a completely saturated model. A RMSEA value less than .06 indicates close approximate fit of the data (Bentler, 1999), whereas a value which approaches 0 indicates perfect fit. In addition, the root mean square error of approximation (RMSEA) adjusts for model complexity so that it favors a simpler model.

The standardized residuals are the discrepancy between the observed correlations and the model-reproduced correlations. An emphasis on standardized residuals can round out the other fit indexes as residual output is not based upon the chi-square test or the size of the correlations (Bentler, 2006). The average of the absolute values of the discrepancy can be considered to be a valid estimation of error. If the smallest discrepancy between observed and reproduced correlations for a variable is about .10, the model is only

marginally wrong, but a discrepancy of .40 is evidence that the model is not fitting the data for the variable in question (Bentler).

Categorical Data Analysis

Generally, variables with a small number of categories are not normally distributed (Kline, 2005). Most items in this study had fewer than three categories. In order to prevent inflated chi-square estimates and the underestimation of factor loadings, categorical data analysis was used to analyze the math scores in the Series 2 models. In order to run categorical data analysis, it was assumed that the exam data was ordinal in nature and that a continuous variable was underlying each categorical variable, which had a scale that was continuous and normal (Bentler, 2006). All ordinal categories are assumed to be ranked in such a way that interpretation of magnitude between categories was consistent.

In the categorical data analyses, the correlation matrix was used for tests of model fit in place of the covariance structure. Polychoric correlations were used as estimates of what the correlations would be if two categories with two or more levels were continuous and normally distributed. Since the categorical variables were not continuous, normal theory estimates based upon univariate and joint statistics had little meaning (Bentler, 2006). Thus, absolute standardized residuals, robust and scaled chi-square statistics were used for the analysis of model fit in the Series 2 models. Measures of multivariate kurtosis were not reported for these models due to the nonnormality of the data.

Improving Model Fit

The primary task in model testing is to determine the goodness of fit between the sample data and the researcher's model (Byrne, 1994). The hypothesized model is a

restricted form of the observed data, and thus perfect fit is highly unlikely (Byrne). The Lagrange Multiplier (LM) is a modification index analogous to forward selection in regression, which predicts the improvement in overall fit if one or more paths were to be added to a model. LM modification indices indicate whether added paths would have a significant impact on goodness of fit measures by indicating the probability of a decrease in the chi-square statistic. The goal of decreasing the chi-square statistic is to be able to get a non-significant result such that a researcher cannot reject the null hypothesis that there is no difference in fit between the observed data and the data as structured by the hypothesized model. The chi-squared goodness of fit tests reported in this paper were evaluated for statistical significance at the $\alpha = .05$ level.

Results

Descriptive Statistics

As seen in Table 2, a correlation of .89 was found between the English and Spanish math scores for the sample of 8th grade examinees. The finding of a statistically significant correlation did not indicate anything about the relative difficulty of the exams, but did indicate that the students ranked similarly on both exams.

An examination of descriptive characteristics of the Spanish and English versions of the math exam revealed statistically significant intercorrelations at the $\alpha = .01$ level for all benchmarks (see Tables A1-A4 in Appendix A). However, an analysis of correlations between item scores in the Spanish and English math tests showed low correlations between items (see Tables A5-A12 in Appendix A). The correlations between item scores ranged from insignificant to a moderate correlation size of .52. Less than half of the items had statistically significant correlations with the other items

measuring the four Algebra benchmarks. An evaluation of the descriptive statistics for item scores showed that item number 29 in the Algebra benchmark, modeling relationships, was not functioning like the other discrete variables. This item,

Table 2. Correlations, Covariances, Standard Deviations

and Means of Sample English and Math Total Raw Scores

English Math Raw Score	Spanish Math Raw Score
	185.78
0.89**	
29.03	26.42
14.06	14.83

***Correlations are significant at the .01 level (2-tailed)*

with a mean score of .01 and a standard deviation of .17 in English, and a mean score of .01 and a standard deviation of .12 in Spanish was removed from the analysis due to the lack of information provided by this variable.

Items 50 and 72 of the Algebra patterns and functions benchmark had the highest percentage of missing data in the English exam, with approximately 24% of the data imputed using EM imputation. In the Spanish version of the exam, item number 50 of the Algebra patterns and functions benchmark, had 30% of the scores imputed.

Series 1. English Benchmark Model

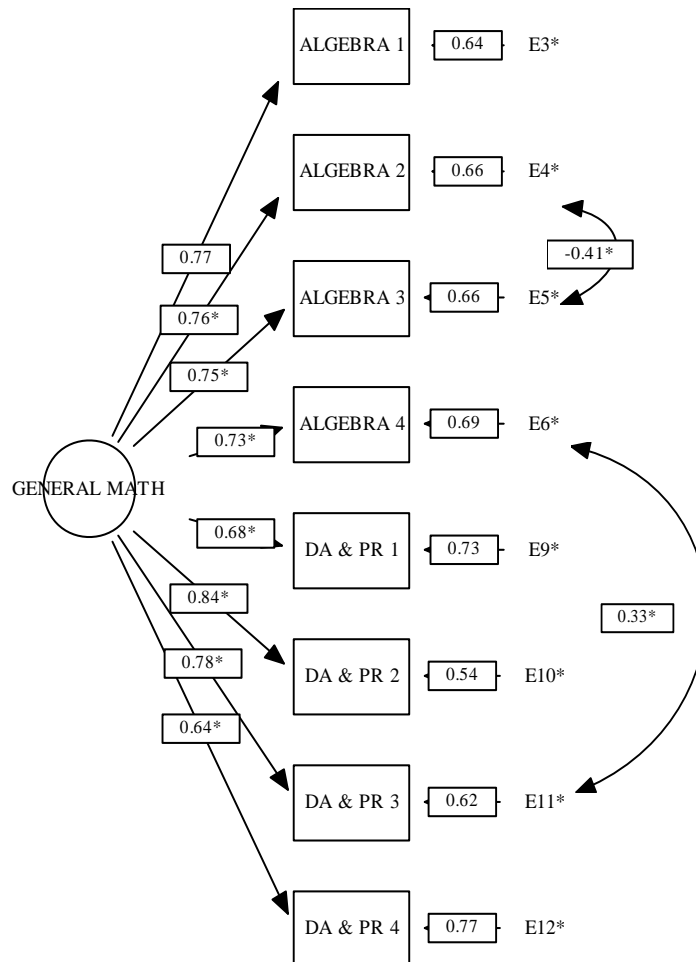
The English two-factor model (see Figure C1 in Appendix C) had an estimated correlation of 1.0 between the factors, which was evidence for a lack of discriminant validity between the standards of Algebra and Data Analysis & Probability. When a correlation between two factors is higher than .85, the two factors cannot be determined

to measure distinct constructs (Kline, 2005). Thus, the two factors were collapsed into a one-factor model with all eight benchmarks as indicator variables. An examination of univariate skew and kurtosis measures did not reveal values that were indicative of nonnormality. Values ranged from -.02 to 1.39 for skewness, and -.85 to 1.67 for kurtosis. Mardia's normalized estimate of .10 did not indicate any problem with multivariate kurtosis. The distribution of standardized residuals looked like a normal distribution with a peak. The removal of cases identified as having the most influence on multivariate kurtosis did not show significant improvement in goodness of fit measures and thus no cases were removed from the data.

The goodness of fit statistics for the one-factor model yielded a CFI of .94, which indicated acceptable fit. However, the chi-square statistic was statistically significant, $\chi^2(20, N = 67) = 36.30, p = .014$. In addition, the RMSEA value of .11 also indicated poor fit to the model. Thus, the Lagrange Multiplier test was applied in order to assess the possibility of improving the model's fit. Two correlated error paths were suggested by the LM test. The assumption of local independence of errors assumes that the performance on one benchmark is not correlated with the performance on another benchmark. However, the addition of the two paths correlating the benchmark errors made sense substantively. It is possible that in addition to measuring overlapping skills, some items measure distinctly different skills such that math performance on one benchmark may be expected to vary with math performance on another benchmark. Due to confidentiality measures, the item questions were not available. Thus, it was not possible to determine the relationship in content between the items measuring the two benchmarks.

The one-factor model was respecified with two paths added as suggested by the LM test. The model was specified using maximum likelihood estimation in five iterations. Since there were no known issues with normality, normal theory statistics were used as indices of model fit. The largest absolute standardized residual in the modified model was .10, which is an acceptable discrepancy (Kline, 2005). The RMSEA measurement index showed improvement at .06, and the CFI value of .98, also indicated improvement in fit (Bentler, 1999). After adding the two paths, the chi-square statistic decreased and was not significant, $\chi^2(18, N = 67) = 22.89, p = .19$. Thus, adding the two paths improved the model's ability to express the variance in test scores. The final model with the standardized parameter estimates is presented in Figure 2.

Figure 2. English one-factor benchmark model



As seen in Figure 2, the path between the Algebra benchmarks of algebraic symbols and modeling relationships had a significant correlation of $-.41$. The negative correlation indicated that, after the common factor variance had been removed, improvement in scores on one benchmark was moderately correlated with the decrease of scores on the other benchmark. In addition, the path between the fourth Algebra benchmark, analyzing changes, and the third Data Analysis & Probability benchmark, data analysis, had a significant moderate correlation of $.33$. The positive correlation indicated that, after the common factor variance had been removed, improvement in scores on one benchmark was moderately correlated with improvement in scores on the second benchmark. The specification of correlations between error terms reflects the assumption that the underlying factor structure is not measuring all of the shared sources of variability in the model (Kline, 2005). The presence of an unmeasured variable was suggested by the significant correlations found between the errors, however, the variable was not known as far as the model was concerned (Kline).

Standardized Estimates. The measurement equations for each estimated parameter represented the causal effect of the factor on the indicator. The standardized estimates seen in Figure 2 indicated that all variables were predicted well by the construct of General Math Knowledge. The standardized parameter estimates represented the effects of General Math Knowledge on each of the benchmarks in the following manner: for one standardized unit increase in the estimate of General Math Knowledge, there was a $.84$ standardized unit increase in student performance on the Data Analysis & Probability benchmark, statistical methods. The equation for the benchmark, statistical methods, had the largest standardized estimate and as expected generated the highest R-

*Table 3. Unstandardized Factor Loadings with Standard Errors
for English One-Factor Benchmark Model*

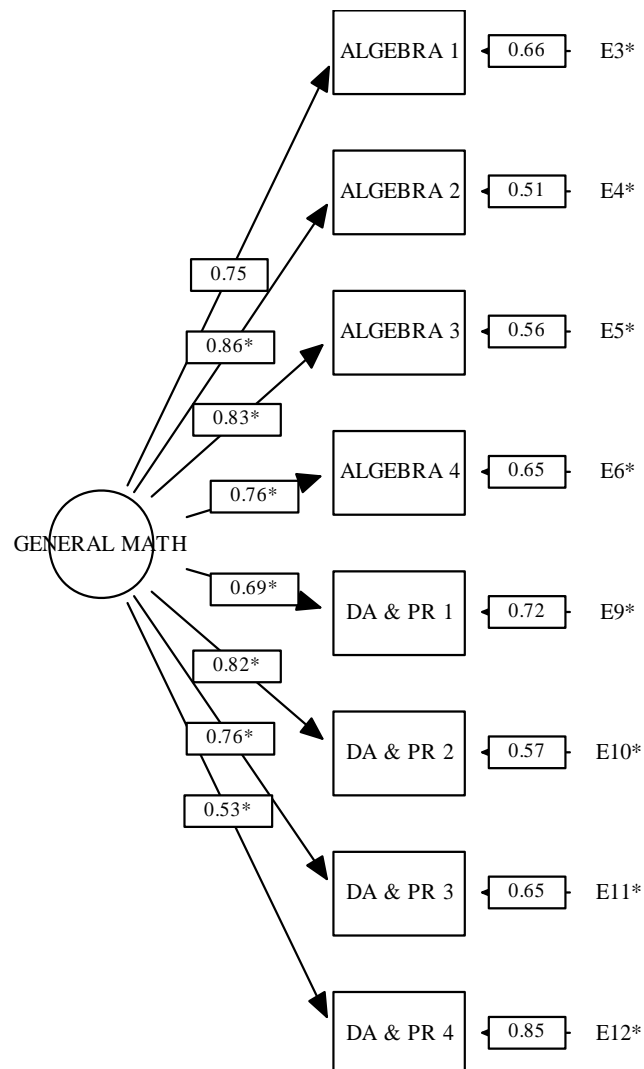
Measure and variable	Unstandardized Factor Loading	SE	p
Algebra Benchmarks			
Patterns & Functions	1.000		
Algebraic Symbols	0.977	0.16	0.000
Modeling Relationships	1.169	0.22	0.000
Analyzing Changes	1.059	0.219	0.000
Data Analysis & Probability Benchmarks			
Data Collection & Representation	0.725	0.095	0.000
Statistical Methods	1.255	0.185	0.000
Data Analysis	1.517	0.258	0.000
Probability	0.587	0.129	0.000
Covariances			
Algebraic Symbols & Modeling Relationships	-0.481	0.147	0.003
Analyzing Changes & Data Analysis	0.536	0.289	0.016

squared value of .71. Thus, 71% of explained variance in the estimate of statistical methods could be attributed to the factor, General Math Knowledge. The standardized parameters, or correlations, between the indicators and the factor were all greater than .60, which gives strong evidence for convergent validity (Kline, 2005). As seen in Table 3, the unstandardized factor loadings including the covariances between independent variables were all statistically significant at the alpha = .01 level. Factor loadings, standard errors, standardized measurement error estimates and R-squared values for the English one-factor model can be found in Table C1 of Appendix C.

Series 1. Spanish Benchmark Model

A correlation of .97 was found between the factors of Algebra and Data Analysis & Probability in the Spanish two-factor model of the math exam (see Figure C3 in Appendix C). Correlations higher than .85 between factors cannot be assumed to measure distinct domains, (Kline, 2005). Thus, the two-factor model was collapsed into a model with one-factor representing the scores on all eight benchmarks. The one-factor model had excellent fit and is presented below in Figure 3 along with the standardized parameters.

Figure 3. Spanish one-factor benchmark model



An examination of univariate skew and kurtosis measures did not indicate any problems with nonnormality. Values ranged from .00 to 1.22 for skewness, and -.40 to 1.53 for kurtosis. In addition, Mardia's normalized estimate of -.91 did not indicate a departure from multivariate normality. The distribution of standardized residuals looked like a normal distribution with a peak. Maximum likelihood estimation as used to run the model, which converged to a solution in six iterations. Normal theory statistics were used as indices of model fit. The RMSEA value of 0.00 and the CFI value of 1.00 indicated excellent fit. The chi-square statistic was not significant, $\chi^2(20, N = 67) = 18.6, p = .55$. The largest absolute correlation residual between variables was .09, which was an indication that the observed correlations could be explained by the hypothesized model correlations.

Standardized Estimates. The lowest standardized estimate was .53 for the probability benchmark, which means that for a one standardized unit increase in the estimate of General Math Knowledge, there was a .53 standardized unit increase in student performance in probability. All other correlations or standardized parameter estimates were greater or equal to .69, with the highest correlation of .86 found for the Algebra benchmark, algebraic symbols.

The term, structural coefficients, is given to parameter estimates that load on a single factor without correlations of errors. When the structural coefficients are statistically significant, the R-squared values for the freely estimated parameters are statistically significant values of explained variance. The largest R-squared value of .74

*Table 4. Unstandardized Factor Loadings with Standard Errors
for Spanish One-Factor Benchmark Model*

Measure and variable	Unstandardized Factor Loading	SE	p
Algebra Benchmarks			
Patterns & Functions	1.000		
Algebraic Symbols	1.049	0.151	0.000
Modeling Relationships	1.349	0.237	0.000
Analyzing Changes	1.166	0.197	0.000
Data Analysis & Probability Benchmarks			
Data Collection & Representation	0.691	0.113	0.000
Statistical Methods	1.304	0.239	0.000
Data Analysis	1.501	0.295	0.000
Probability	0.347	0.073	0.000

was contributed by the equation for the second Algebra benchmark, algebraic symbols. Thus, 74% of the variance in the estimate of student performance on algebraic symbols could be attributed to the factor of General Math Knowledge.

Kline (2005) suggests that parameters with loadings greater or equal to .60 provide evidence for convergent validity. Relatively high loadings on math achievement are to be expected if the indicators are measures of the same theoretical construct. In contrast, loadings smaller than .20 would indicate that additional factors might be needed in the model (Kline). Although the standardized parameter estimate for the probability benchmark was .53, the R-squared value of .28 was statistically significant and showed that the factor was an important predictor of explained variance in the probability

measurement equation. As seen in Table 4, all of the unstandardized coefficients were statistically significant at the $\alpha = .01$ level. Factor loadings, standard errors, standardized measurement error estimates and R-squared values for the Spanish one-factor model can be found in Table C2 of Appendix C.

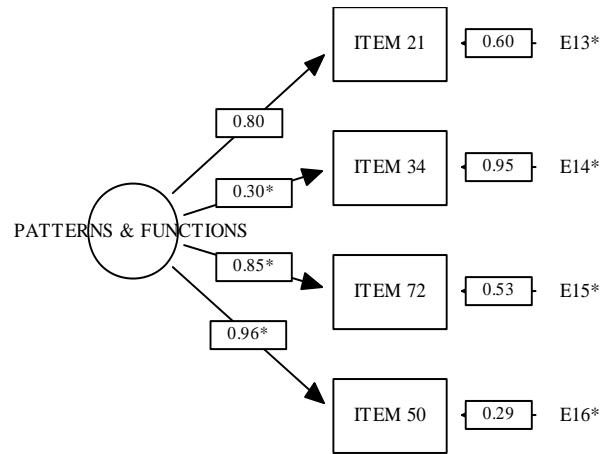
Series 2. English Benchmark, Patterns and Functions

An English one-factor model was run with item numbers 21, 34, 50 & 72 as indicators measuring the construct, Patterns & Functions. Categorical analysis with ML estimation of the correlation matrices was utilized. No special problems were encountered during optimization and the model was identified in six iterations. The standardized residual matrix reported an absolute value of .14 between items 34 and 72. While this value is greater than the cut-off value of .10 recommended by Kline (2005), it is still not large enough to warrant concern about the discrepancy between the sample polychoric correlations and the model polychoric correlations. The distribution of standardized residuals looked like a normal distribution with a peak. The goodness of fit robust statistics showed a non-significant Satorra-Bentler scaled chi-square statistic, $\Pi^2(2, N = 67) = 1.70, p = .43$, which indicated good fit. The CFI value of 1.00 and the RMSEA value of 0.00 showed excellent results of model fit. The model along with the standardized parameter estimates is presented in Figure 4.

Data Analysis. The parameter of item 21 was set to a fixed value of 1.00 and of the three remaining freely estimated parameters, only the loadings on items 50 & 72 were statistically significant at the .05 level. The smallest loading of .30 was on item 34, which meant that for one standardized unit increase in the estimate of the Patterns and Functions factor, there was only a .30 standardized unit increase in Item 21. However,

the remaining three items showed high loadings greater or equal to .80.

Figure 4. English pattern and function item level model



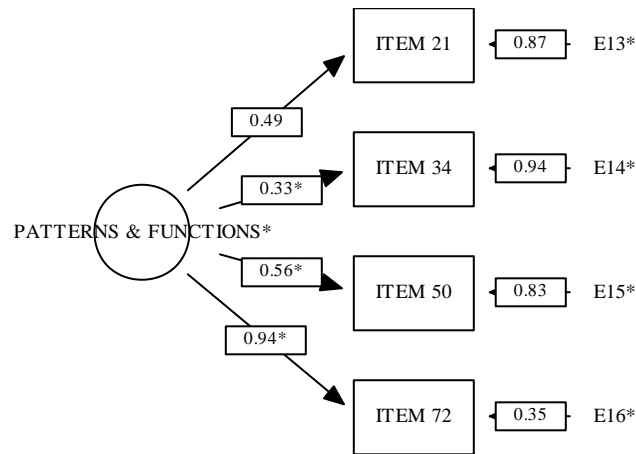
Because the items loaded on one factor, the corresponding R-squared values were equal to the squared standardized coefficients. As expected, the measurement equation for item 34 resulted in the lowest R-squared contribution of .09. Only 9% of the variance in student performance on item 34 could be accounted for by the factor, Patterns and Functions. Evidence for convergent validity was not seen in this model due to the low loading on item 34. See Table C3 in Appendix C for factor loadings, standard errors, standardized measurement error estimates and R-squared values.

Series 2. Spanish Benchmark, Patterns and Functions

A Spanish one-factor model was run with item numbers 21, 34, 50 & 72 as observed variables measuring the construct, Patterns & Functions. The model converged on a solution in 18 iterations. The distribution of standardized residuals looked normal. An examination of the standardized residual matrix revealed the largest absolute residual of .15 between items 21 and 34. A standardized residual greater than .10 might be of concern, but excessive departure from fit between the sample and model polychoric

correlations was not seen. The Satorra-Bentler scaled chi-square statistic was non-significant, $\Pi^2(2, N = 67) = 2.75, p = .25$, which indicated good fit. The RMSEA value of .08 indicated marginal fit (Kline, 2005), but was not unreasonable given the categorical nature of the data. The CFI value of .96 indicated good fit. The model along with the standardized parameter estimates is presented in Figure 5.

Figure 5. Spanish pattern and function item level model



Data Analysis. The parameter of item 21 was fixed to a value of 1.00 for scaling purposes. Although item 72 had a standardized parameter of .94 (with a R-squared value of .88), the corresponding standard error of 1.29 was high relative to the unstandardized estimate of 1.91. Thus, item 72 had a non-significant test statistic of 1.49 at the $\alpha = .05$ level. With a statistically significant unstandardized estimate, item 50 had the next highest R-squared value of .32. Thus, 32% of the variance in student performance on item 50 was explained by the construct, Patterns and Functions, in the Spanish model. The smallest loading of .33 was seen for item 34. Although evidence for convergent validity was not consistent, the correlations between the factor loadings and the construct were not lower than .20. The presence of factor loadings smaller than .20 suggests that

an indicator did not correlate substantially with the latent variable (Kline, 2005). See Table C3 in Appendix C for factor loadings, standard errors, standardized measurement error estimates and R-squared values.

Series 2. English Benchmark, Algebraic Symbols

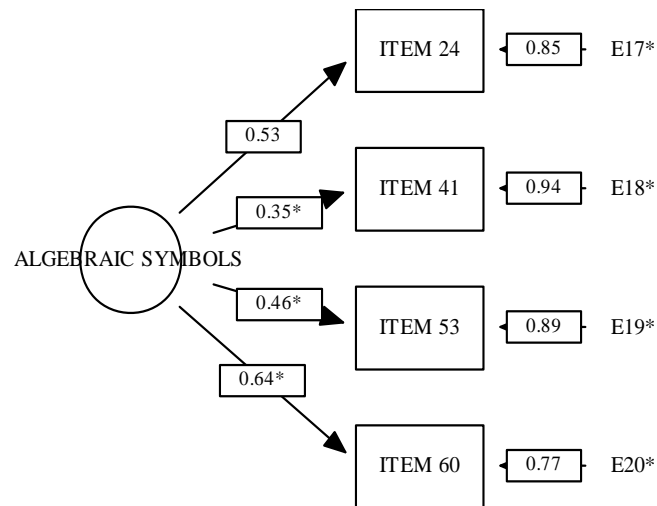
Items 24, 41, 53 and 60 were the indicators of the one-factor English model, Algebraic Symbols. Categorical data analysis with robust and scaled chi-square statistics were utilized due to the discrete nature of the variables. No special problems were encountered during optimization and the model was fit in six iterations using maximum likelihood estimation. The intercorrelations between variables were small with the highest correlation of .34 found between items 24 and 60 (see Table A6 in Appendix A). The distribution of standardized residuals looked normal and the largest absolute standardized residual was .06, which did not give cause for concern regarding the fit between the sample polychoric correlations and the model polychoric correlations.

The Satorra-Bentler scaled chi-square statistic was non-significant, $\Pi^2(2, N = 67) = .25, p = .88$, which indicated good fit. The RMSEA value of 0.00 and the CFI index of 1.00 were both indicators of excellent fit. The model along with the standardized parameter estimates is presented in Figure 6.

The parameter for item 24 was fixed to 1. Utilizing robust statistics, none of the remaining three parameters were found to be statistically significant at the $\alpha = .05$ level. Item 41 had the lowest standardized estimate of .35, whereas the highest standardized estimate of .64 was found for item 60. The R-squared values for each parameter were equal to the square of the standardized estimates such that the lowest and highest R-squared values of .12 and .41 were seen for item 41 and item 60 respectively.

Approximately 41% of the variance in student performance on item 60 was explained

Figure 6. English algebraic symbols item level model



by the factor, Algebraic Symbols. The range of standardized estimates from .35 to .64 indicated a lack of evidence for convergent validity (see Table C3 in Appendix C for factor loadings, standard errors, standardized measurement error estimates and R-squared values).

Conclusion

Confirmatory factor analysis is distinct from exploratory factor analysis in that it allows for the testing of hypothesized models that have been grounded in empirical research (Byrne, 1994). Thus, CFA was well suited for the task of analyzing the internal structure of the NM 8th grade state math assessment in English and Spanish. The goodness of fit indices were within acceptable range for all models evaluated in this study. However, in standard CFA, one-factor models that have three or more indicators are likely to be identified (Kline, 2005). Thus, the results of the models in this study should be viewed cautiously.

Examination of the benchmark structure of the standards of Algebra and Data Analysis & Probability did not reveal evidence for discriminant validity due to high correlations between the standards (Kline, 2005). Thus, the two standards were not found to be distinct measures of mathematical knowledge. The lack of evidence for discriminant validity suggested that the math standards of Algebra and Data Analysis & Probability measured overlapping math skills and knowledge. The expectation for measures of math knowledge to be distinct from one another might not be realistic as student performance may be expected to be consistent within subject areas. However, the possibility that math skills were being tested too broadly should be investigated in order to determine whether the test was measuring constructs of interest adequately.

The analysis of the English and Spanish one-factor models revealed strong evidence for convergent validity. All of the benchmarks in both versions of the exams were moderately-to-highly correlated with the latent factor of General Math Knowledge. The results in Series 1 of this study showed consistency between the benchmark correlations and the factor of General Math Knowledge. Although this study did not assess the mean differences between the English and Spanish versions of the 8th grade exam, pronounced differences in correlations between the English and Spanish models were not noted.

The models in Series 2 were specified in order to evaluate the correlations between items and each of the four Algebra benchmarks. As noted in the methods section, five of the eight Algebra benchmark English and Spanish models had inadmissible solutions due to the likelihood of Heywood cases. A generalization concerning the functioning of the Algebra standard at the item level was not attainable,

thus the remaining three item level models were evaluated primarily to help with the purpose of establishing a methodology for establishing internal structural validity of exam scores. Evidence for convergent validity in these models was high for some benchmarks and low for others. Interpretation of the results was difficult because of the lack of consistency in correlations on factor loadings. Many of the parameter estimates were not statistically significant even with robust standard errors. As expected, the R-squared values for the measurement equations in the item level models tended to be small and insignificant.

One possibility for the lack of evidence for convergent validity at the item score level was the need for a larger sample size. In addition, it was likely that low correlations between the items contributed to the inadmissible model solutions. The sample selected in this study was reduced by the number of students not taking the exam in both languages, as well as, missing data on item scores. In addition, initial editing of the data revealed that item questions and item scoring were not always equivalent between English and Spanish versions of the exam, which led to the removal of data from the analyses. The interpretability of the models in Series 1 as compared to the models in Series 2 was due in part to issues with measurement error. The aggregation of items scores in Series 1 resulted in a smoothing effect that was not seen in Series 2, where the issue of nonnormality with categorical data was predominant. An additional limitation of this study was that language scores were not available, which meant that the covariate of language ability could not inform an evaluation of the difference in means between the English and Spanish versions of the exam.

Recommendations for future research would include the evaluation of group

means, as well as, a study about whether item level difficulty might be another factor accounting for the lack of consistency in the size of item correlations in the Series 2 models. An important purpose of securing evidence for internal structural validity is to support a correct interpretation of test scores (AERA, APA & NCME, 1999). It is hoped that this study has assisted in the interpretation of test scores relative to the internal structure of the 8th grade math exam. In addition, one of the main purposes of this study was to illustrate a methodology for assessing the validity of test exams. Hopefully, the problems encountered in this study can inform future research in the area of test measurement.

References

- American Educational Research Association, American Psychological Association, and National Council on Measurement in Education. (1999). *Standards for Educational and Psychological Testing*. Washington, DC: AERA.
- Bentler, P.M. (1985 – 2006). EQS for windows (Build 83) [computer program]. Encino, CA: Multivariate Software, Inc.
- Bentler, P.M. (2006). *EQS 6 Structural Equations Program Manual*. Encino, CA: Multivariate Software, Inc.
- Byrne, B. M. (1994). *Structural Equation Modeling with EQS and EQS/Windows*. Thousand Oaks, CA. Sage Publications.
- Chen, F., Bollen, K.A., Paxton, P., Curran, P.J., & Kirby, J.B. (2001). Improper solutions in structural equation models: Causes, consequences, and strategies. *Sociological Methods and Research*, 29, 468-508. Retrieved November 24, 2007, from Article First database.
- Hu, L., & Bentler, P. M. (1999). Cutoff Criteria for Fit Indexes in Covariance Structure Analysis: Conventional Criteria versus New Alternatives. *Structural Equation Modeling*. 6 (1), 1-55. Retrieved November 17, 2007 from Academic Search Complete database.
- Kline, R. B., (2005). *Principles and Practice of Structural Equation Modeling*. 2nd Edition. New York: The Guilford Press.
- La Marca, Paul M. (2001). Alignment of standards and assessments as an accountability criterion. *Practical Assessment, Research & Evaluation*, 7921). (retrieved April 29, 2007 from ERIC Document Reproduction Service, ERIC No. ED 458 288).

New Mexico Statewide Assessment Program (NMSAP) 2006-2007 Procedures Manual.

Published by the New Mexico Public Education Department Assessment and Accountability Division. Santa Fe, NM.

No Child Left Behind Act of 2001, Pub. L. No. 107-110, 115 Stat. 14-25 (2002).

SPSS 13.0 for Windows (1989-2006). [Computer software]. SPSS Incorporated.

Stevens, J., Farley, M., & Zvoch, M. (2006). Examination of the internal structural validity of the terranova – CTBS/5. Paper presented at the American Research Association Meeting, San Francisco, CA.

Stansfield, C. W. (2003). Test translation and adaptation in public education in the USA. *Language Testing*, 2, 189-207. Retrieved April 27, 2007 from Academic Search Complete database.

List of Tables in Paper

Table 1. 8th Grade Math Standards and Benchmarks

Table 2. Correlations, Covariances, Standard Deviations and Means of Sample English
and Math Total Raw Scores

Table 3. Unstandardized Factor Loadings with Standard Errors for English One-Factor
Benchmark Model

Table 3. Unstandardized Factor Loadings with Standard Errors for Spanish One-Factor
Benchmark Model

List of Figures in Paper

Figure 1. Sample confirmatory factor analysis model diagram

Figure 2. English one-factor benchmark model

Figure 3. Spanish one-factor benchmark model

Figure 4. English pattern and function item level model

Figure 5. Spanish pattern and function item level model

Figure 6. English algebraic symbols item level model

Appendix A

Table A1. Correlations, Covariances, Standard Deviations and Means of English Algebra Benchmarks

Table A2. Correlations, Covariances, Standard Deviations and Means of English Data Analysis & Probability Benchmarks

Table A3. Correlations, Covariances, Standard Deviations and Means of Spanish Algebra Benchmarks

Table A4. Correlations, Covariances, Standard Deviations and Means of Spanish Data Analysis & Probability Benchmarks.

Table A5. Correlations, Covariances, Standard Deviations and Means of English Item Level Scores for the Patterns and Functions Benchmark

Table A6. Correlations, Covariances, Standard Deviations and Means of English Item Level Scores for the Algebraic Symbols Benchmark

Table A7. Correlations, Covariances, Standard Deviations and Means of English Item Level Scores for the Modeling Relationships Benchmark

Table A8. Correlations, Covariances, Standard Deviations and Means of English Item Level Scores for the Analyzing Changes Benchmark

Table A9. Correlations, Covariances, Standard Deviations and Means of Spanish Item Level Scores for the Patterns and Functions Benchmark

Table A10. Correlations, Covariances, Standard Deviations and Means of Spanish Item Level Scores for the Algebraic Symbols Benchmark

Table A11. Correlations, Covariances, Standard Deviations and Means of Spanish Item Level Scores for the Modeling Relationships Benchmark

Table A12. Correlations, Covariances, Standard Deviations and Means of Spanish Item Level Scores for the Analyzing Changes Benchmark

Table A13. Percentages of Missing Data for Algebra Benchmark, Patterns and Functions, in English and Spanish Versions of the Exam.

Table A14. Percentages of Missing Data for Algebra Benchmark, Algebraic Symbols, in English and Spanish Versions of the Exam.

Table A15. Percentages of Missing Data for Algebra Benchmark, Modeling Relationships, in English and Spanish Versions of the Exam.

Table A16. Percentages of Missing Data for Algebra Benchmark, Analyzing Changes, in English and Spanish Versions of the Exam.

Table A1. Correlations, Covariances, Standard Deviations and

Means of English Algebra Benchmarks

2A (Patterns & Functions)	2B (Algebraic Symbols)	2C (Modeling Relationships)	2D (Analyzing Changes)
	1.39	1.54	1.19
0.61**		1.08	1.44
0.56**	0.39**		1.92
0.46**	0.56**	0.62**	
1.52	1.51	1.81	1.70
1.55	2.43	3.55	2.66

**Correlations are significant at the .01 level (2-tailed)

Table A2. Correlations, Covariances, Standard Deviations and

Means of English Data Analysis & Probability Benchmarks

5A (Data Collection & Representation)	5B (Statistical Methods)	5C (Data Analysis)	5D (Probability)
	1.12	1.50	0.67
0.52**		2.56	1.07
0.54**	0.65**		1.43
0.51**	0.57**	0.59**	
1.24	1.75	2.26	1.08
1.79	1.99	2.73	1.90

**Correlations are significant at the .01 level (2-tailed)

Table A3. Correlations, Covariances, Standard Deviations and

Means of Spanish Algebra Benchmarks

2A (Patterns & Functions)	2B (Algebraic Symbols)	2C (Modeling Relationships)	2D (Analyzing Changes)
	1.47	1.69	1.37
0.71**		1.77	1.55
0.62**	0.70**		2.07
0.53**	0.65**	0.66**	
1.50	1.38	1.83	1.72
1.64	1.67	3.45	3.09

**Correlations are significant at the .01 level (2-tailed)

Table A4. Correlations, Covariances, Standard Deviations and

Means of Spanish Data Analysis & Probability Benchmarks

5A (Data Collection & Representation)	5B (Statistical Methods)	5C (Data Analysis)	5D (Probability)
	1.20	1.36	0.32
0.59**		2.50	0.64
0.54**	0.63**		0.62
0.38**	0.48**	0.38**	
1.13	1.79	2.23	0.74
1.39	2.73	2.16	1.00

**Correlations are significant at the .01 level (2-tailed)

Table A5. Correlations, Covariances, Standard Deviations and Means

of English Item Level Scores for the Patterns and Functions Benchmark

Item Number 21	Item Number 34	Item Number 50	Item Number 72
	0.05	0.16	0.14
0.18		0.07	0.03
0.50**	0.26*		0.15
0.52**	0.13	0.57	
0.58	0.48	0.57	0.48
0.40	0.36	0.34	0.16

**Correlations are significant at the .01 level (2-tailed)

*Correlations are significant at the .05 level (2-tailed)

Table A6. Correlations, Covariances, Standard Deviations and Means

of English Item Level Scores for the Algebraic Symbols Benchmark

Item Number 24	Item Number 41	Item Number 53	Item Number 60
	0.06	0.03	0.05
0.17		0.04	0.05
0.12	0.12		0.05
0.21	0.13	0.20	
0.45	0.74	0.50	0.49
0.27	0.69	0.57	0.39

Table A7. Correlations, Covariances, Standard Deviations and Means

of English Item Level Scores for the Modeling Relationships Benchmark

Item Number 29	Item Number 37	Item Number 40	Item Number 61	Item Number 69
	0.03	-0.01	0.02	0.00
0.15		0.11	0.08	0.17
-0.14	0.22		0.02	0.02
0.23	0.11	0.10		0.07
0.01	0.28*	0.09	0.23	
0.17	1.19	0.57	0.50	0.40
0.03	2.00	0.27	0.48	0.81

*Correlations are significant at the .05 level (2-tailed)

Table A8. Correlations, Covariances, Standard Deviations and Means

of English Item Level Scores for the Analyzing Changes Benchmark

Item Number 39	Item Number 43	Item Number 67	Item Number 75
	0.01	0.07	0.04
0.04		0.17	0.04
0.15	0.38**		0.16
0.15	0.18	0.37**	
0.49	0.50	0.88	0.50
0.60	0.58	0.72	0.46

**Correlations are significant at the .01 level (2-tailed)

Table A9. Correlations, Covariances, Standard Deviations and Means

of Spanish Item Level Scores for the Patterns and Functions Benchmark

Item Number 21	Item Number 34	Item Number 50	Item Number 72
	0.05	0.01	0.10
0.20		0.32	0.14
0.03	0.17		0.40
0.32**	0.20	0.40**	
0.48	0.50	0.57	0.64
0.36	0.49	0.27	0.48

***Correlations are significant at the .01 level (2-tailed)*

Table A10. Correlations, Covariances, Standard Deviations and Means

of Spanish Item Level Scores for the Algebraic Symbols Benchmark

Item Number 24	Item Number 41	Item Number 53	Item Number 60
	0.10	0.03	0.02
0.30*		0.03	0.10
0.12	0.07		0.00
0.07	0.30*	-0.01	
0.46	0.75	0.50	0.46
0.30	0.66	0.51	0.30

**Correlations are significant at the .05 level (2-tailed)*

Table A11. Correlations, Covariances, Standard Deviations and Means

of Spanish Item Level Scores for the Modeling Relationships Benchmark

Item Number 29	Item Number 37	Item Number 40	Item Number 61	Item Number 69
	0.01	0.01	0.00	-0.01
0.06		0.17	0.23	0.06
0.12	0.31**		0.09	0.03
-0.05	0.42**	0.36**		0.08
-0.09	0.12	0.11	0.32**	
0.12	1.08	0.50	0.51	0.48
0.01	2.48	0.49	0.21	0.34

**Correlations are significant at the .01 level (2-tailed)

Table A12. Correlations, Covariances, Standard Deviations and Means

of Spanish Item Level Scores for the Analyzing Changes Benchmark

Item Number 39	Item Number 43	Item Number 67	Item Number 75
	0.08	0.06	-0.04
0.31*		0.11	0.01
0.16	0.32**		0.08
-0.18	0.03	0.23	
0.50	0.50	0.72	0.50
0.58	0.54	0.57	0.43

*Correlations are significant at the .05 level (2-tailed)

**Correlations are significant at the .01 level (2-tailed)

*Table A13. Percentages of Missing Data for Algebra Benchmark,
Patterns and Functions, in English and Spanish Versions of the Exam*

Test Version	Item Number 21	Item Number 34	Item Number 50	Item Number 72
English	1.49	0.00	23.88	23.88
Spanish	10.45	0.00	29.85	17.91

*Table A14. Percentages of Missing Data for Algebra Benchmark,
Algebraic Symbols, in English and Spanish Versions of the Exam*

Test Version	Item Number 24	Item Number 41	Item Number 53	Item Number 60
English	0	1.49	0	0
Spanish	0.00	14.93	0.00	4.48

*Table A15. Percentages of Missing Data for Algebra Benchmark,
Modeling Relationships, in English and Spanish Versions of the Exam*

Test Version	Item Number 29	Item Number 37	Item Number 40	Item Number 61	Item Number 69
English	8.96	1.49	0.00	7.46	0.00
Spanish	31.34	1.49	0.00	16.42	7.46

*Table A16. Percentages of Missing Data for Algebra Benchmark,
Analyzing Changes, in English and Spanish Versions of the Exam*

Test Version	Item Number 39	Item Number 43	Item Number 67	Item Number 75
English	0.00	0.00	11.94	7.46
Spanish	0.00	1.49	13.43	10.45

Appendix B

Table B1. Skewness and Kurtosis Measures for English Series 1 Benchmark Variables

Table B2. Skewness and Kurtosis Measures for Spanish Series 1 Benchmark Variables

Figure B1. English algebra benchmark, patterns and functions, scatter plot

Figure B2. English algebra benchmark, algebraic symbols, scatter plot

Figure B3. English algebra benchmark, modeling relationships, scatter plot

Figure B4. English algebra benchmark, analyzing changes, scatter plot

Figure B5. Spanish algebra benchmark, patterns and functions, scatter plot

Figure B6. Spanish algebra benchmark, algebraic symbols, scatter plot

Figure B7. Spanish algebra benchmark, modeling relationships, scatter plot

Figure B8. Spanish algebra benchmark, analyzing changes, scatter plot

Figure B9. English data analysis and probability benchmark, data collection and
representation, scatter plot

Figure B10. English data analysis and probability benchmark, statistical methods, scatter
plot

Figure B11. English data analysis and probability benchmark, data analysis, scatter plot

Figure B12. English data analysis and probability benchmark, probability, scatter plot

Figure B13. Spanish data analysis and probability benchmark, data collection and
representation, scatter plot

Figure B14. Spanish data analysis and probability benchmark, statistical methods, scatter
plot

Figure B15. Spanish data analysis and probability benchmark, data analysis, scatter plot

Figure B16. Spanish data analysis and probability benchmark, probability, scatter plot

Figure B17. English algebra benchmark, patterns and functions, histogram

Figure B18. English algebra benchmark, algebraic symbols, histogram

Figure B19. English algebra benchmark, modeling relationships, histogram

Figure B20. English algebra benchmark, analyzing changes, histogram

Figure B21. Spanish algebra benchmark, patterns and functions, histogram

Figure B22. Spanish algebra benchmark, algebraic symbols, histogram

Figure B23. Spanish algebra benchmark, modeling relationships, histogram

Figure B24. Spanish algebra benchmark, analyzing changes, histogram

Figure B25. English data analysis and probability benchmark, data collection and
representation, histogram

Figure B26. English data analysis and probability benchmark, statistical methods,
histogram

Figure B27. English data analysis and probability benchmark, data analysis, histogram

Figure B28. English data analysis and probability benchmark, probability, histogram

Figure B29. Spanish data analysis and probability benchmark, data collection and
representation, histogram

Figure B30. Spanish data analysis and probability benchmark, statistical methods,
histogram

Figure B31. Spanish data analysis and probability benchmark, data analysis, histogram

Figure B32. Spanish data analysis and probability benchmark, probability, histogram

*Table B1. Skewness and Kurtosis Measures for**English Series 1 Benchmark Variables*

	<u>Skewness</u>	<u>Kurtosis</u>
<u>Algebra Factor</u>		
Patterns & Functions	1.385	1.66
Algebraic Symbols	0.782	0.263
Modeling Relationships	-0.023	-0.671
Analyzing Changes	0.231	-0.826
<u>Data Analysis & Probability Factor</u>		
Data Collection & Representation	0.499	-0.421
Statistical Methods	0.937	0.092
Data Analysis	0.631	-0.85
Probability	0.209	-0.707

*Table B2. Skewness and Kurtosis Measures for**Spanish Series 1 Benchmark Variables*

	<u>Skewness</u>	<u>Kurtosis</u>
<u>Algebra Factor</u>		
Patterns & Functions	1.218	1.175
Algebraic Symbols	0.888	0.138
Modeling Relationships	0.099	-0.17
Analyzing Changes	0.22	-0.404
<u>Data Analysis & Probability Factor</u>		
Data Collection & Representation	1.048	1.529
Statistical Methods	0.602	-0.033
Data Analysis	0.705	-0.73
Probability	0	-1.139

Figure B1. English algebra benchmark,
patterns and functions, scatter plot

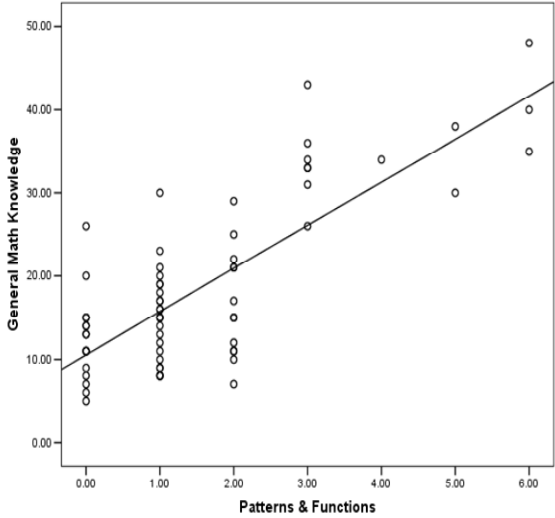


Figure B2. English algebra benchmark,
algebraic symbols, scatter plot

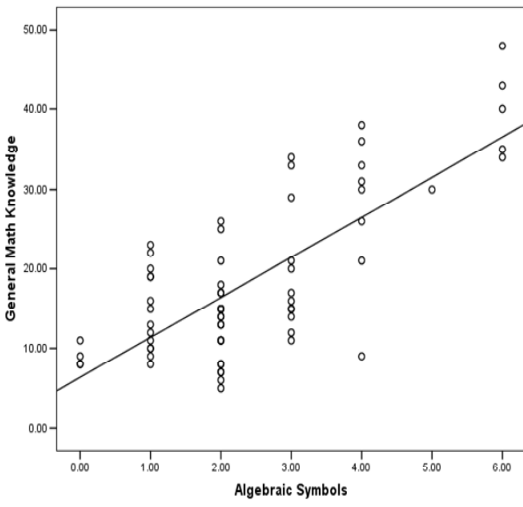


Figure B3. English algebra benchmark,
modeling relationships, scatter plot

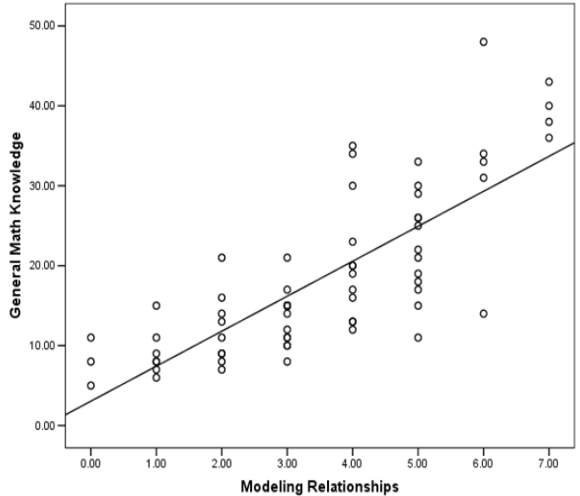


Figure B4. English algebra benchmark,
analyzing changes, scatter plot

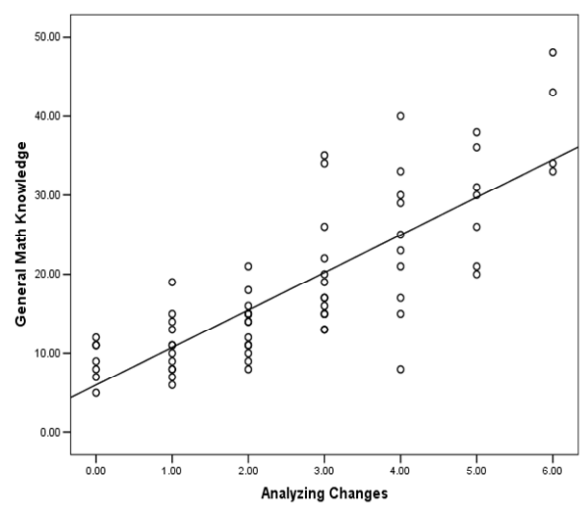


Figure B5. Spanish algebra benchmark,

patterns and functions, scatter plot

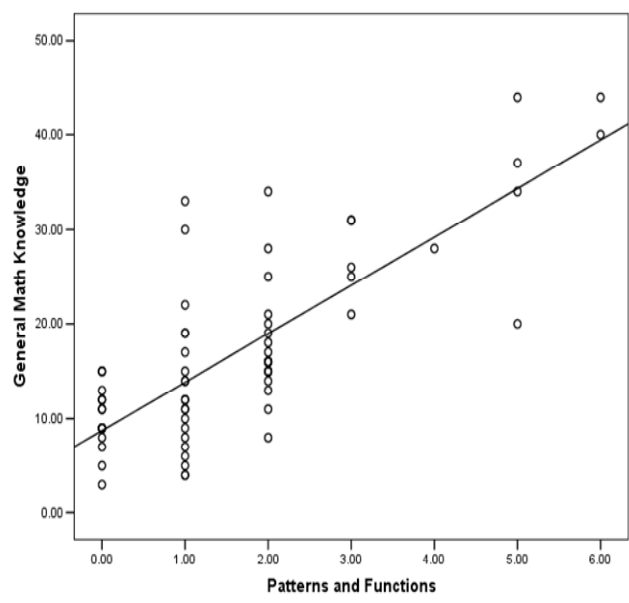


Figure B7. Spanish algebra benchmark,

modeling relationships, scatter plot

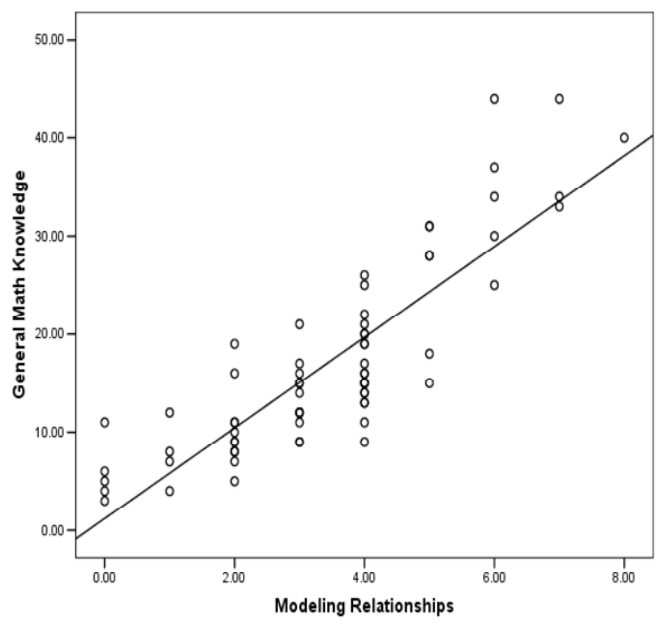


Figure B6. Spanish algebra benchmark,

algebraic symbols, scatter plot

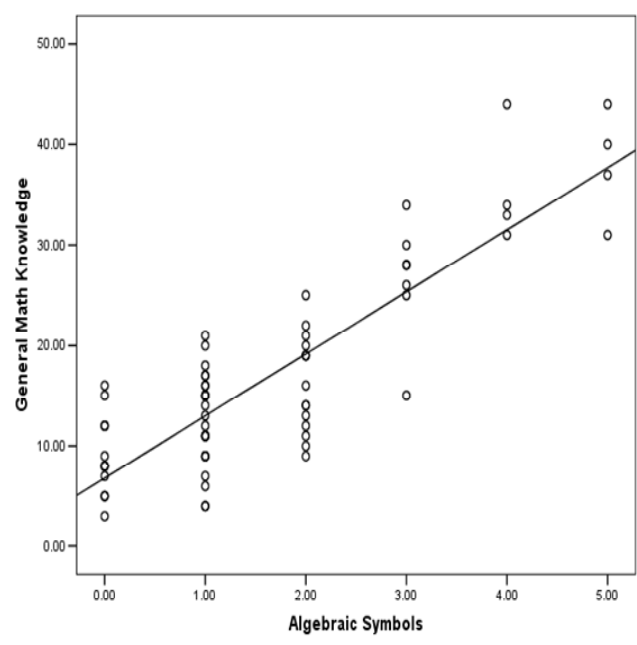


Figure B8. Spanish algebra benchmark,

analyzing changes, scatter plot

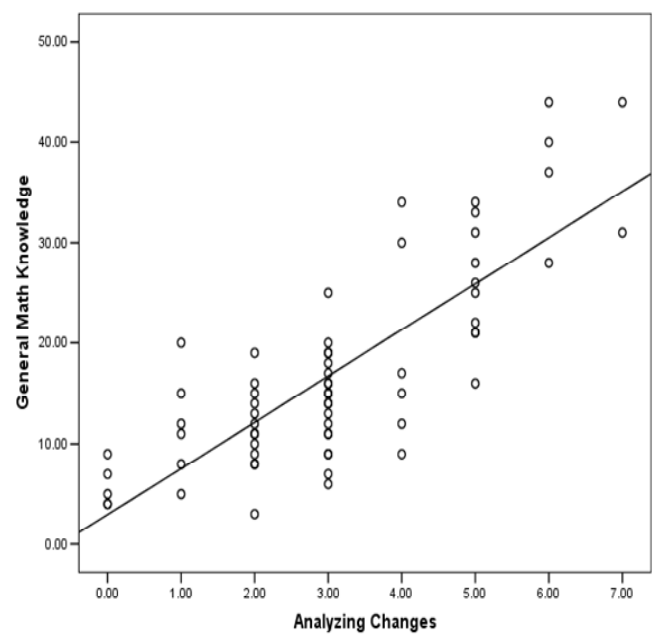


Figure B9. English data analysis and probability

benchmark, data collection and representation, scatter plot

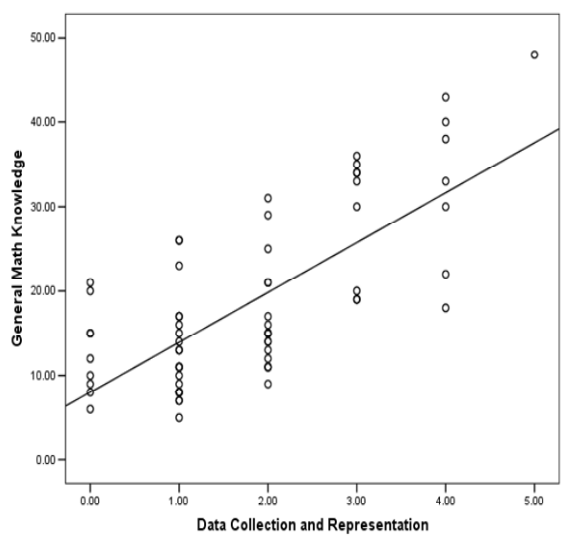


Figure B11. English data analysis and probability

benchmark, data analysis, scatter plot

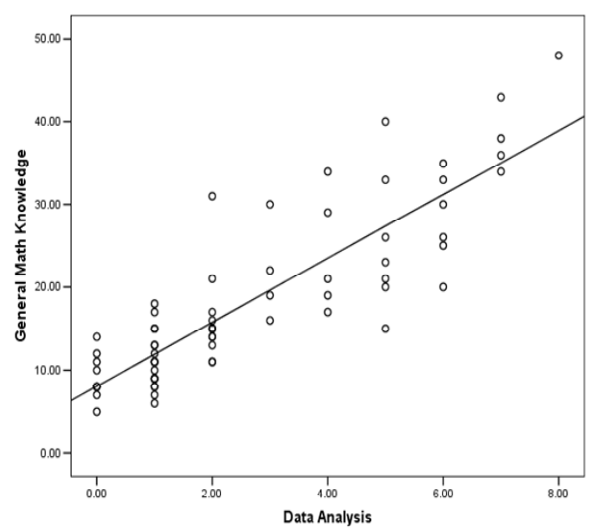


Figure B10. English data analysis and probability

benchmark, statistical methods, scatter plot

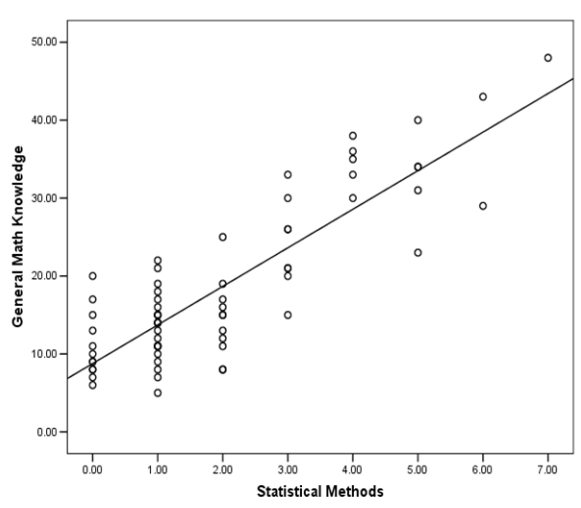


Figure B12. English data analysis and probability

benchmark, probability, scatter plot

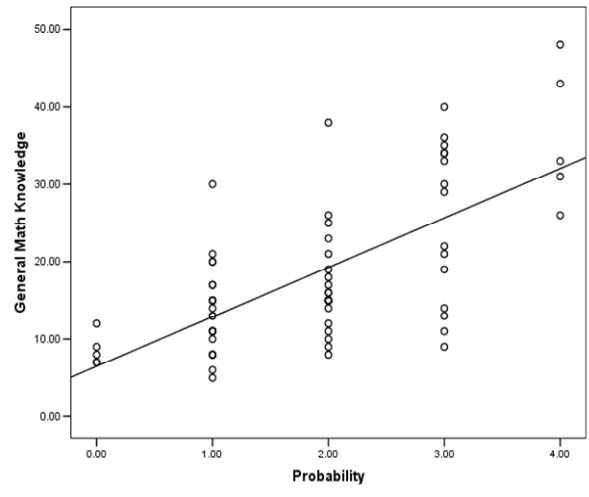


Figure B13. Spanish data analysis and probability

benchmark, data collection and representation, scatter plot

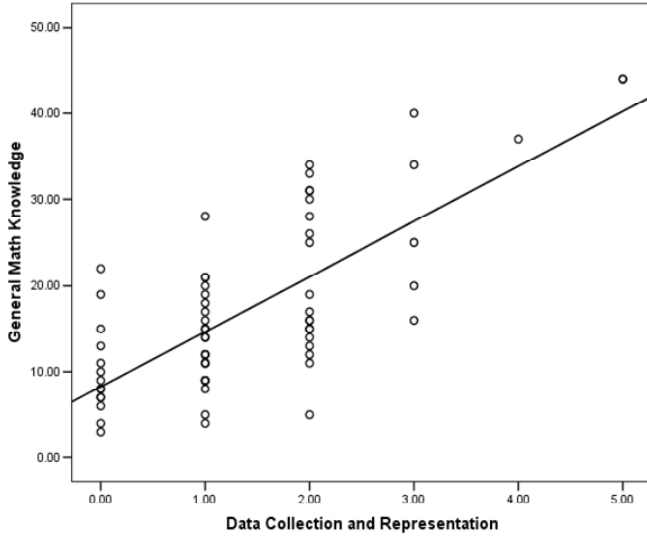


Figure B15. Spanish data analysis and probability

benchmark, data analysis, scatter plot

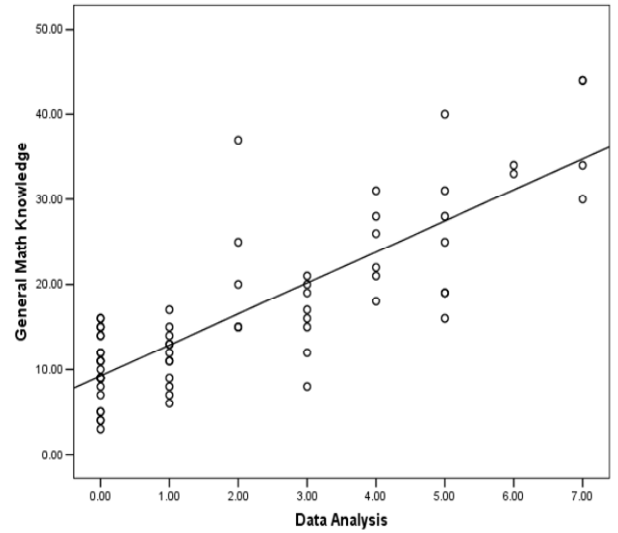


Figure B14. Spanish data analysis and probability

benchmark, statistical methods, scatter plot

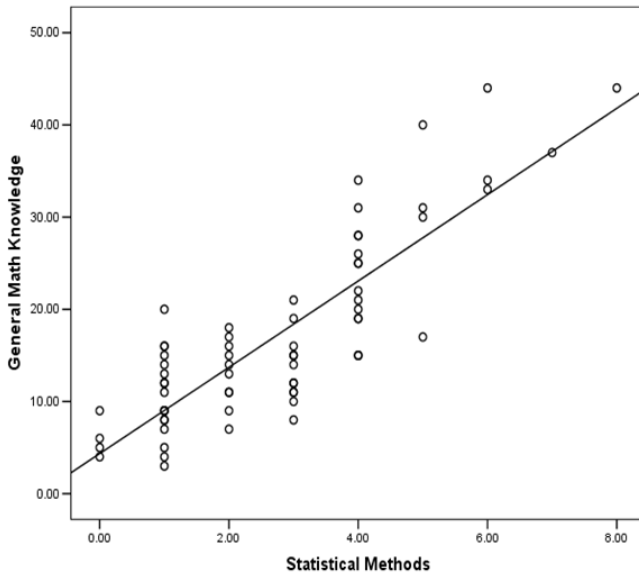


Figure B16. Spanish data analysis and probability

benchmark, probability, scatter plot

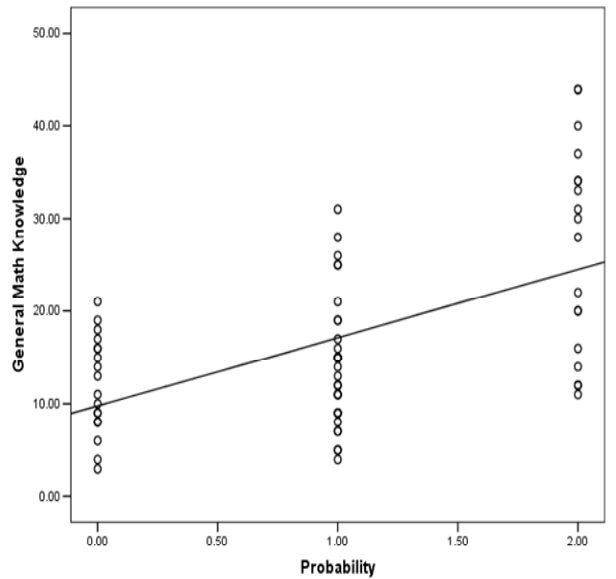


Figure B17. English algebra benchmark,
patterns and functions, histogram

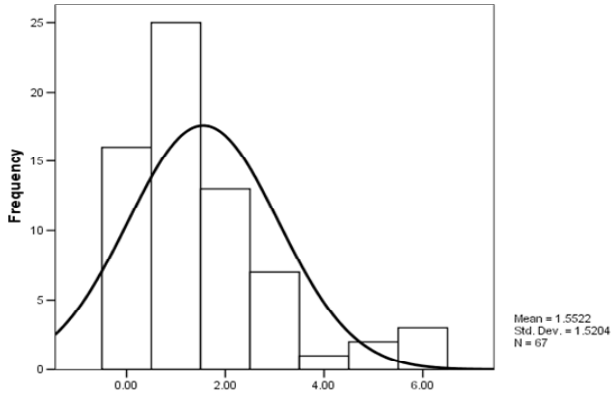


Figure B18. English algebra benchmark,
algebraic symbols, histogram

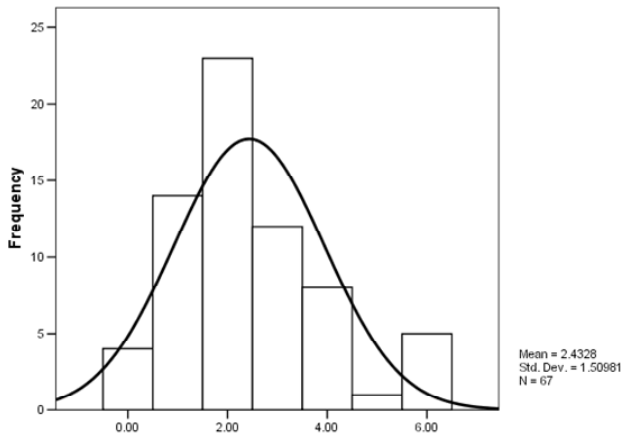


Figure B19. English algebra benchmark,
modeling relationships, histogram

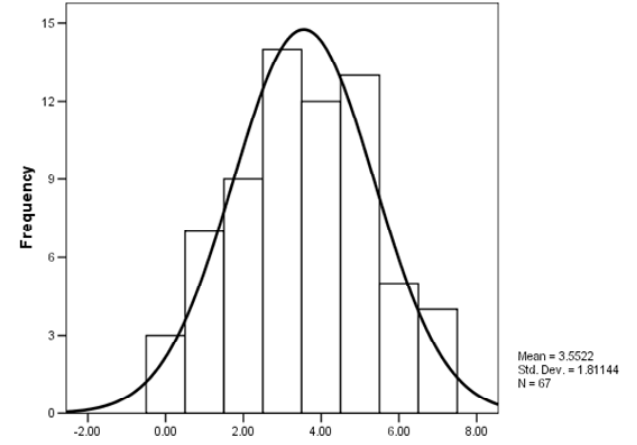


Figure B20. English algebra benchmark,
analyzing changes, histogram

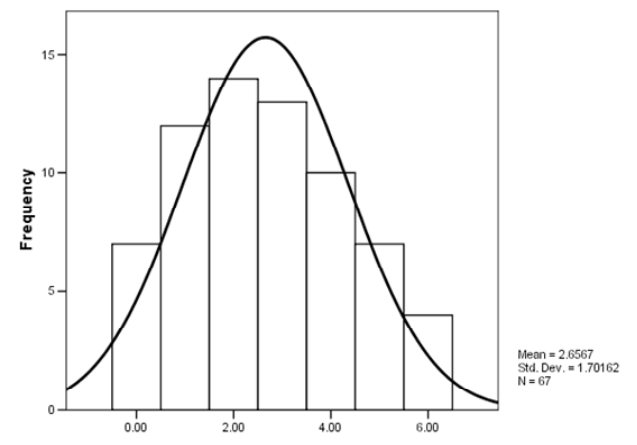


Figure B21. Spanish algebra benchmark,
patterns and functions, histogram

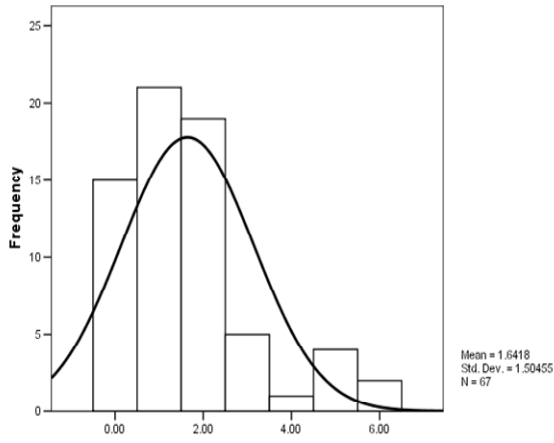


Figure B22. Spanish algebra benchmark,
algebraic symbols, histogram

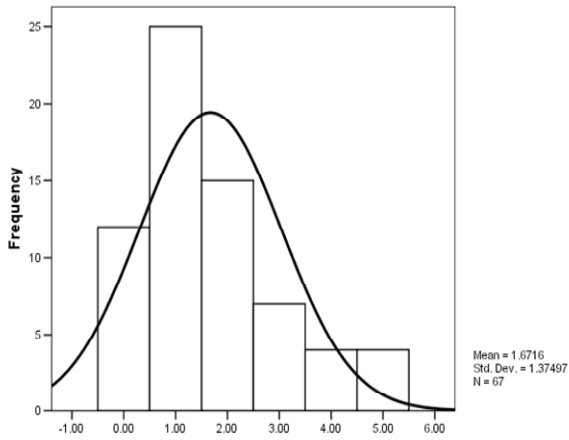


Figure B23. Spanish algebra benchmark,
modeling relationships, histogram

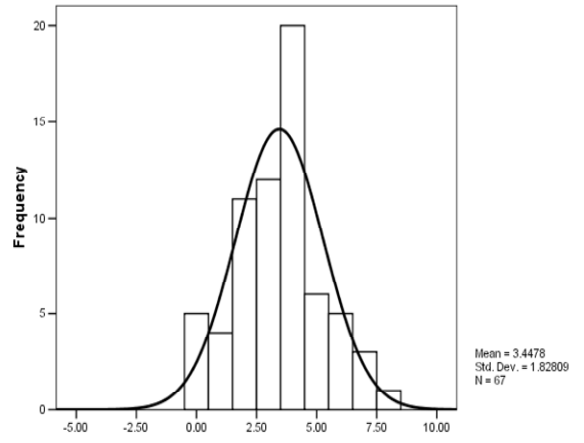


Figure B24. spanish algebra benchmark,
analyzing changes, histogram

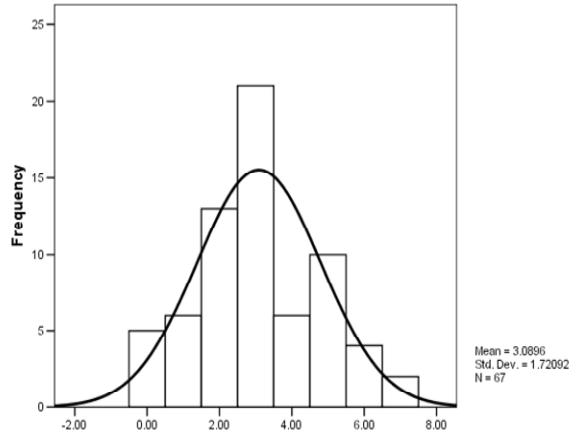


Figure B25. English data analysis and probability

benchmark, data collection and representation, histogram

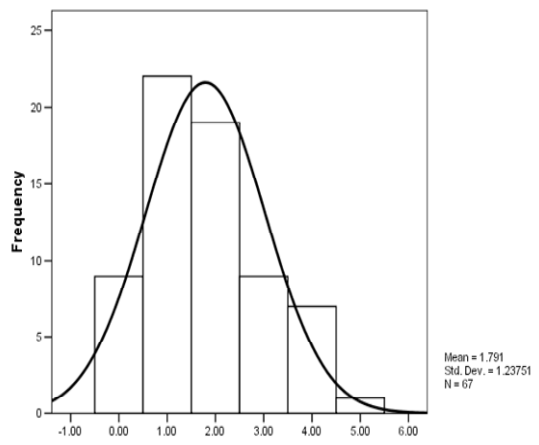


Figure B26. English data analysis and probability

benchmark, statistical methods, histogram

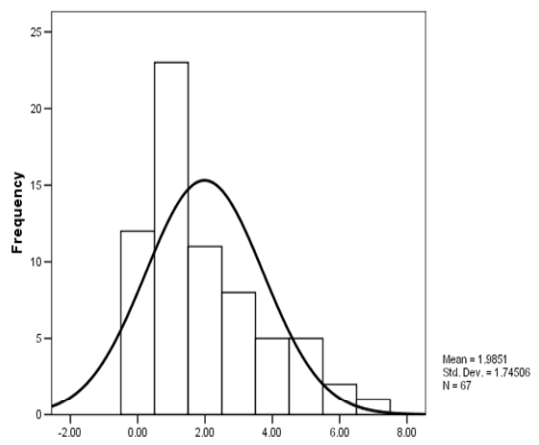


Figure B27. English data analysis and probability

benchmark, data analysis, histogram

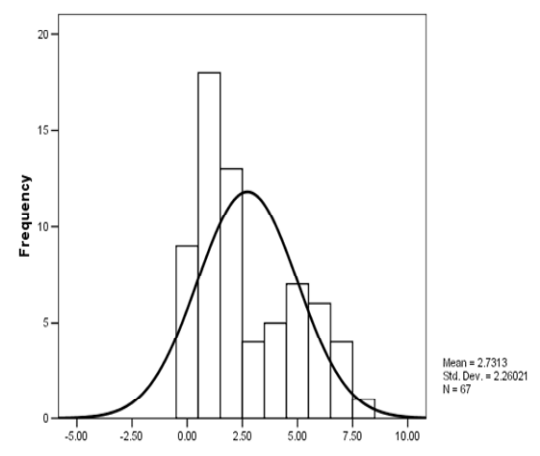


Figure B28. English data analysis and probability

benchmark, probability, histogram

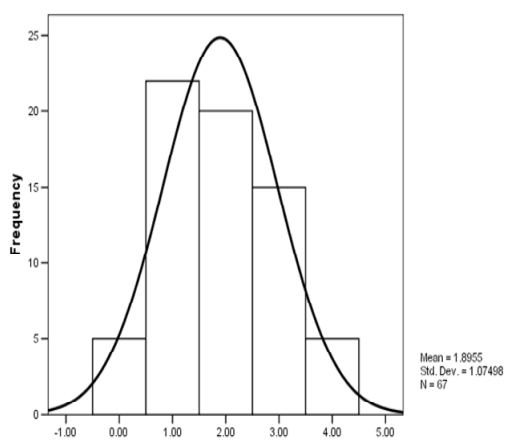


Figure B29. Spanish data analysis and probability

benchmark, data collection and representation, histogram

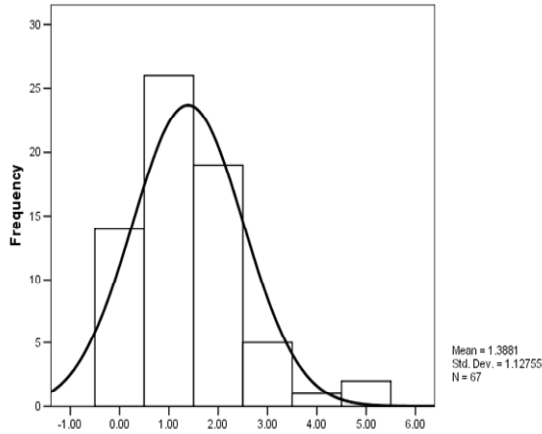


Figure B31. Spanish data analysis and probability

benchmark, data analysis, histogram

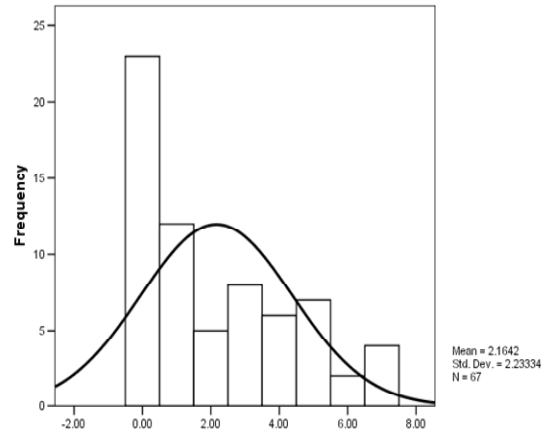


Figure B30. Spanish data analysis and probability

benchmark, statistical methods, histogram

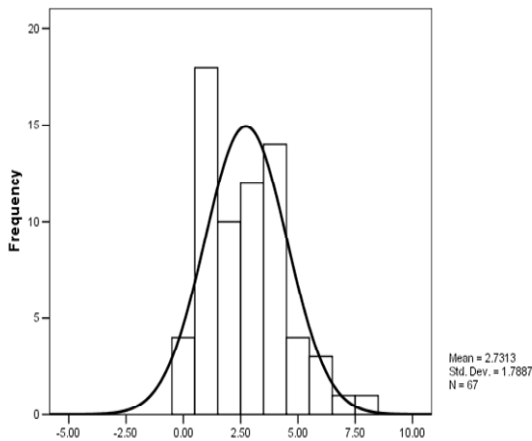
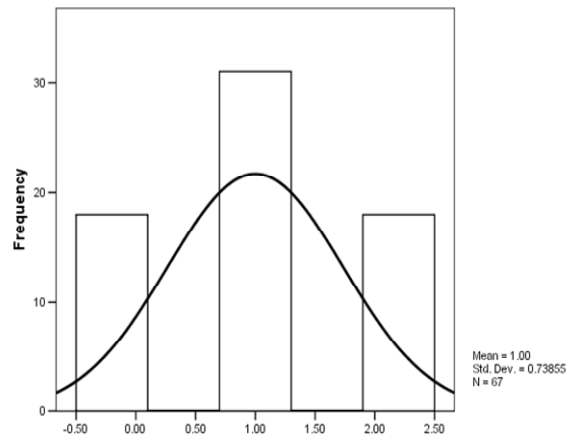


Figure B32. Spanish data analysis and probability

benchmark, probability, histogram



Appendix C

Table C1. Factor Loadings and Uniqueness for Confirmatory Factor Model of English

Benchmark Variables

Table C2. Factor Loadings and Uniqueness for Confirmatory Factor Model of Spanish

Benchmark Variables

Table C3. Factor Loadings and Uniqueness for Confirmatory Factor Model of Item Level

Variables with Robust Standard Errors

Figure C1. English benchmark two-factor path diagram

Figure C2. English benchmark one-factor path diagram

Figure C3. Spanish benchmark two-factor path diagram

Figure C4. Spanish benchmark one-factor path diagram

Figure C5. Item level path diagram for english algebra benchmark, patterns and functions

Figure C6. Item level path diagram for spanish algebra benchmark, patterns and functions

Figure C7. Item level path diagram for english algebra benchmark, algebraic symbols

Table C1. Factor Loadings and Uniqueness for Confirmatory Factor Model
of English Benchmark Variables

Measure and variable	Unstandardized Factor Loading	SE	Standardized Factor Loading	Uniqueness	R^2
Algebra Factor -- Benchmarks					
Patterns & Functions	1.000	1.000	0.768	0.641	0.589
Algebraic Symbols	0.977	0.155	0.755	0.655	0.571
Modeling Relationships	1.169	0.186	0.753	0.658	0.567
Analyzing Changes	1.059	0.174	0.727	0.687	0.528
Data Analysis & Probability Factor -- Benchmarks					
Data Collection & Representation	0.725	0.127	0.683	0.73	0.467
Statistical Methods	1.255	0.173	0.84	0.543	0.705
Data Analysis	1.517	0.227	0.783	0.622	0.614
Probability	0.587	0.111	0.637	0.770	0.406
Covariances					
Algebraic Symbols & Modeling Relationships	-0.481	0.175	-0.408	--	--
Analyzing Changes & Data Analysis	0.536	0.25	0.326	--	--
Variance					
Factor	1.362	0.378			

Table C2. Factor Loadings and Uniqueness for Confirmatory Factor Model
of Spanish Benchmark Variables

Measure and variable	Unstandardized Factor Loading	SE	Standardized Factor Loading	Uniqueness	R^2
Algebra Factor -- Benchmarks					
Patterns & Functions	1.000	1.000	0.749	0.663	0.560
Algebraic Symbols	1.049	0.146	0.859	0.512	0.738
Modeling Relationships	1.349	0.194	0.831	0.556	0.691
Analyzing Changes	1.166	0.185	0.763	0.646	0.582
Data Analysis & Probability Factor -- Benchmarks					
Data Collection & Representation	0.691	0.122	0.69	0.724	0.476
Statistical Methods	1.304	0.19	0.821	0.571	0.674
Data Analysis	1.501	0.24	0.757	0.654	0.573
Probability	0.347	0.082	0.529	0.848	0.280
Variance					
Factor	1.268	0.364			

Table C3. Factor Loadings and Uniqueness for Confirmatory Factor Models
of Item Level Variables with Robust Standard Errors

Measure and variable	Unstandardized Factor Loading	SE	Standardized Factor Loading	Uniqueness	R^2
English Factor -- Patterns & Functions					
Item 21	1.000	1.000	0.800	0.601	0.639
Item 34	0.380	0.297	0.304	0.953	0.092
Item 50	1.197	0.228	0.957	0.289	0.916
Item 72	1.06	0.096	0.848	0.531	0.719
Variance					
Factor	0.639	0.131			
English Factor -- Algebraic Symbols					
Item 24	1.000	1.000	0.533	0.846	0.284
Item 41	0.658	0.477	0.351	0.936	0.123
Item 53	0.868	0.630	0.463	0.886	0.214
Item 60	1.196	0.852	0.638	0.770	0.407
Variance					
Factor	0.284	0.28			
Spanish Factor -- Patterns & Functions					
Item 21	1.000	1.000	0.491	0.871	0.241
Item 34	0.678	0.396	0.333	0.943	0.111
Item 50	1.145	0.582	0.562	0.827	0.316
Item 72	1.909	1.289	0.937	0.349	0.878
Variance					
Factor	0.241	0.211			

Figure C1. English benchmark two-factor path diagram

Figure C2. English benchmark one-factor path diagram

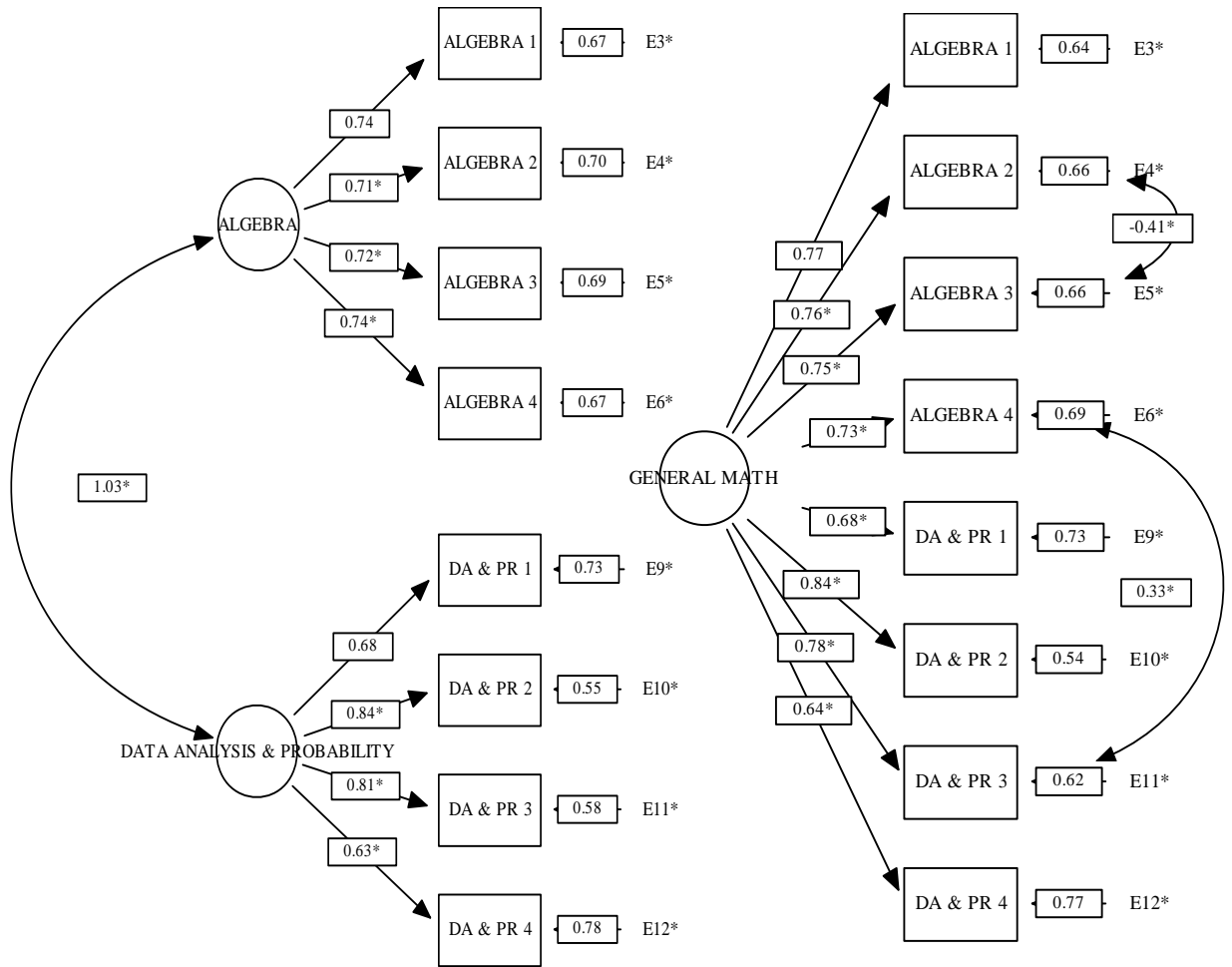


Figure C3. Spanish benchmark two-factor path diagram

Figure C4. Spanish benchmark one-factor path diagram

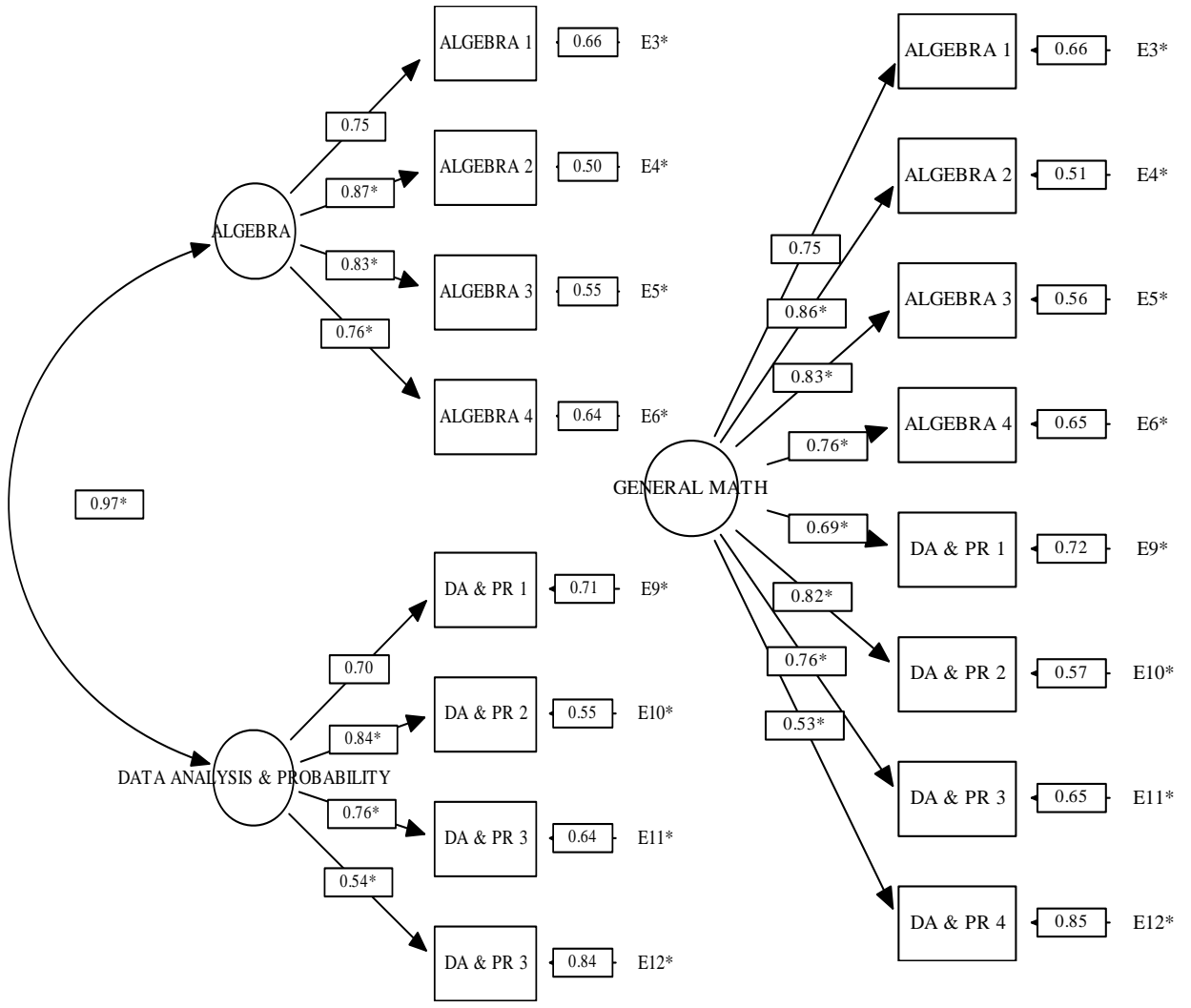


Figure C5. Item level path diagram for english algebra

Figure C7. Item level path diagram for english algebra

benchmark, patterns and functions

benchmark, algebraic symbols

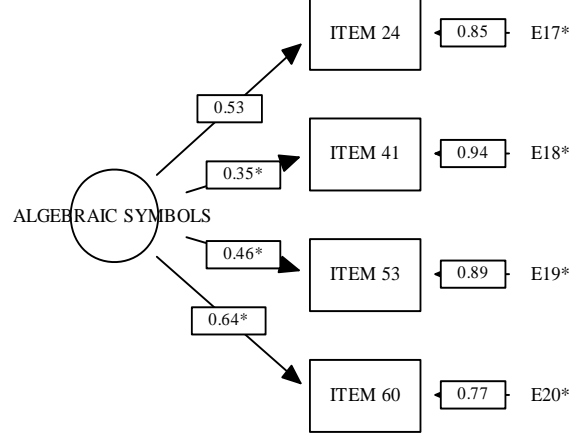
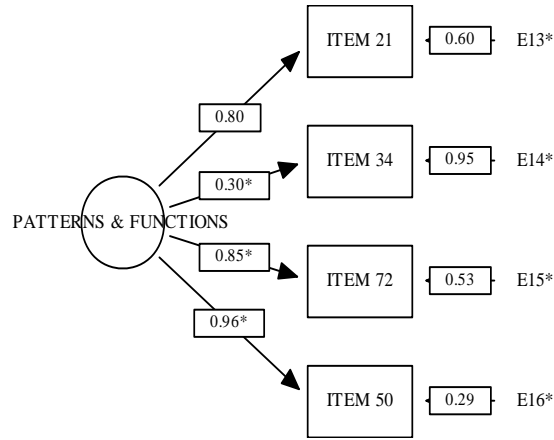


Figure C6. Item level path diagram for spanish algebra

benchmark, patterns and functions

