Edgenuity™

Analysis
of the Reliability and Validity of an Edgenuity Algebra I Quiz

This study presents the steps Edgenuity uses to evaluate the reliability and validity of its quizzes, topic tests, and cumulative exams.

To illustrate the evaluation process, we randomly selected a course and quiz from our curriculum. The selected assessment is a polynomial quiz from the Algebra I course with an "extensive" reliability factor of .75. Six certified teachers and content experts evaluated the quiz's content validity. The item-objective congruence ranged from .76 to 1.00, and interviews indicate strong content-related validity for the quiz. A confirmatory factor analysis found the quiz to have exceptionally strong construct validity: X^2 (72) = 75.23, p = .37; RMSEA = .01(.90 CI = .000 - .029); and CFI = .995.

Introduction

The purpose of this study is to present the procedure Edgenuity uses to evaluate the reliability and validity of its assessments. The following section discusses the rigorous analysis used to establish a measure's reliability, as well as its content and construct validity.

The structural analysis of the polynomial quiz demonstrates a precise alignment of Edgenuity test items to their relevant objectives.

To illustrate the process, we began by randomly selecting an assessment from a course within the Edgenuity curriculum. Using a Monte Carlo model to generate random numbers, we selected the first quiz from the 10th lesson of the Algebra I course. This polynomial quiz had nine questions aligned to the objective (describe polynomials) and six questions aligned to a second objective (add and subtract polynomials). A total of 15 questions were evaluated on the quiz. An example item is shown below:

Simplify $(x^4 + 2x^2 - 5x) + (-3x^3 + x^2 + 1) + (3x^4 + 2x)$

A $3x^8 - 2x^4 + 3x^3 - 3x^2 + 1$

B $4x^4 - 3x^3 + x^2 + 4x + 2$

C $4x^4 - 3x^3 + 3x^2 + 4x + 2$

D $4x^4 - 3x^3 + 3x^2 - 3x + 1$

Figure 1. Polynomial quiz question aligned to the objective: add and subtract polynomials.

To correctly answer this item, students must first identify like terms, then combine them by performing the indicated operation on their coefficients. Since all terms are being added together, one can get rid of the parenthesis/grouping symbols because no distribution needs to take place. Next, one must add the x^4 to the $3x^4$ to get the lead term $4x^4$; the term that is the second-highest degree, $-3x^3$, is not added to any other term; while $2x^2$ and x^2 are added to get $3x^2$; etc. We finally begin assembling the polynomial from the solutions to form the correct answer, $4x^4 – 3x^3 + 3x^2 – 3x + 1$ (choice D).

To establish this quiz's reliability and validity, we considered the following questions:

- What is the reliability of the polynomial quiz?
- Does the polynomial quiz demonstrate strong content validity?
- Does the quiz have strong structural fit to the empirical data?

Reliability and Cronbach's Alpha

We sought to answer the research question, "What is the reliability of the polynomial quiz?" The quiz was found to have a reliability of .75, considered an extensive reliability.

The best approach for determining the reliability of the polynomial quiz is based on internal, or inter-item, consistency. Internal consistency is the accuracy and trustworthiness of an assessment in measuring an intended goal (Thorndike, 1997). It is a type of reliability concerned with the homogeneity of items within a scale. For

example, suppose the construct being measured was the presence of cancer in the human body and the test used was a blood test. We would want the blood test to be incredibly accurate (reliable) and the amount of blood drawn to be small. We would also want each blood sample to be homogenous, to demonstrate consistency. If half the vials of blood tested indicated the presence of cancer while the other half did not, the test would be considered unreliable (only 50 percent agreement). In our case, the number of samples taken that define the objective (describe polynomials) is nine, analogous to taking nine vials of blood. In essence, this is internal consistency reliability.

"Although we cannot directly observe the linkage between items and the latent variable, we can certainly determine whether the items are correlated to one another" (DeVellis, 2003). Of a number of possible approaches, the Cronbach's alpha measure was chosen to assess the reliability of the polynomial quiz. When the variable being measured is dichotomous, as in our case, the Cronbach's coefficient alpha becomes the Kuder-Richardson formula 20, or KR-20 (Kuder & Richardson, 1937).

Possible values for the KR-20 reliability index fall into the range 0 to 1. Although different tests have different criteria, a widely accepted criterion is that tests with a reliability index higher than .70 are reliable for group measurement, and tests with an index above .80 are reliable for individual measurement (Thorndike, 1997). Robinson, Shaver, and Wrightsman (1991) would consider an alpha of .80 or better to be exemplary, .70 to .79 extensive, .60 to .69 moderate, and less than .60 minimal. Under most circumstances, evaluation instruments are designed to be used with samples of large groups of students (> 200) (Holden, Fekken & Cotton, 1991). Thus, with the combination of our sample size of 465 and a reliability index of .75, one can safely claim the test is reliable.

The polynomial quiz was found to have a reliability of .75, considered an extensive reliability.

Content Validity

We sought to answer the research question, "Does the polynomial quiz demonstrate strong content validity?" The results of the analysis provided strong content-related validity for the polynomial quiz.

Various definitions of content validity have been published in *Standards for Educational and Psychological Testing* (American Educational Research Association, 1985), but most definitions encompass concepts embodied in the following statement: "Content validity is the degree to which elements of an assessment instrument are relevant to and representative of the targeted construct or objective for a particular assessment purpose" (Haynes, Richard, Kubany, 1995).

Content validation should incorporate both quantitative and qualitative methods applicable to all elements of the assessment (Haynes et al., 1995; Thorndike, 1997; DeVellis, 2003). A recommended method of establishing content validity is to have experts review items and establish their relevance to the intended topic (Haynes et al., 1995; Thorndike, 1997; DeVellis, 2003).

To do this, Edgenuity researchers asked content experts familiar with topics covered on the polynomial quiz to judge each question individually as well as the quiz as a whole. In keeping with the multimethod recommendations, items were inspected qualitatively through interviews and quantitatively with a three-point scale using the index of item-objective congruence established by Rovinelli and Hambleton (1977). The process of item-objective congruence requires content experts to rate how well individual items measure specific objectives listed by the test developer.

To ensure the evaluators remained independent, the content experts were not told which constructs the individual items were intended to measure nor that they were related to this study.

More specifically, a content expert evaluates each item by giving it a rating of 1 (clearly measuring), -1 (clearly not measuring), or 0 (unclear degree of measurement) for each objective. For example, a value of -1 for the valid objective would indicate the experts believe the item is not measuring the intended objective, but rather an unintended objective. The premise of the index is to have high positive values on the intended objective and values close to -1 on each remaining objective.

The choice of a cutoff score for this index to separate "good" from "bad" items is based on the Rovinelli and Hambleton (1977) recommendation that three-quarters of the content specialists judge an item to be a perfect match to an objective, while other judges are not able to make a decision.

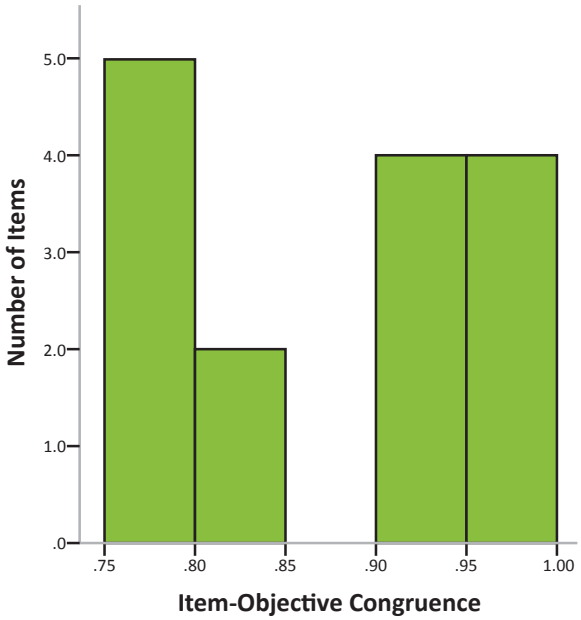

Figure 2. Histogram of item-objective congruence.

Thus, test constructors obtaining an index of .75 would know that at a minimum, at least 75 percent of the content specialists rendered a perfect rating. Six content experts, all certified teachers of mathematics who did not take part in the quiz's construction, were selected to evaluate its validity. The results of the analysis are shown in Figure 2.

We can see that two items had an item-objective congruence between .80 and .85, while eight items had values between .90 and 1.00. Overall results of testing indicated that item indices of congruency ranged from 1.00 to .76 with an average of .86 (*SD* = .10). This range, average, and standard deviation indicate the content experts were in agreement that all 15 items corresponded to their intended objective, demonstrating strong content validity for the polynomial quiz. Further, the interviews confirmed the quantitative results, such that the content writer's descriptions overlapped greatly with each item's intended measurement.

Construct Validity and Confirmatory Factor Analysis

We sought to answer the research question, “Does the quiz have strong structural fit to the empirical data?” The structural analysis of the polynomial quiz demonstrates a precise alignment of Edgenuity test items to their relevant objectives.

Construct validity subsumes all categories of validity and is the degree to which an assessment instrument measures the targeted construct (Messick, 1993). It is the extent to which the empirical data reflect the underlying intended model, based on the alignment of item to objective (DeVellis, 2003). The most widely used method for establishing construct validity is in the use of factor analysis; whenever factor analysis was used over other methods, the results identified subscales even when other analyses did not find any (Clark & Watson, 1995).

Structural equation modeling (SEM) is an extension of factor analysis, and therefore a general linear model and multiple regression analysis procedure (Kline, 1998). SEM is used to determine the degree to which sample data fit a theoretical model hypothesized by a researcher (Schumacker & Lomax, 2004). SEM analysis combines confirmatory factor analysis (CFA) that tests the relationship between a latent variable and the set of observed variables that define it. “The goal of structural equation modeling and confirmatory factor analysis is to determine if the hypothesized theoretical model, which depicts the causal pattern of relationships that was determined a priori, is reflected by the empirical data” (Lei & Wu, 2007).

While the checks may give general information about whether a model has specification errors or may imply that a data set does or does not fit, testing a hypothesis requires measures of good fit that can be tested for significance. This chi-square has degrees of freedom based on the "number of non-redundant variances and covariances of observed variables, and the total number of parameters to be estimated" (Mueller, 1996). Used as a test statistic with a set significance level (typically $p<.05$), a significant chi-square implies that the null hypothesis of a perfect fit of data to model must be rejected (significance is not what we want). Despite a variety of model fit indices in addition to the chi-squared statistic, two of those most widely used for assessments are the comparative fit index (CFI) and the root mean square error of approximation (RMSEA) (Kline, 1998). These two indices do not correct for the number of parameter estimates (Kline, 1998), and tend to be the most interpretable due to established criteria.

Essentially, the CFI and RMSEA indices work in opposite ways. To address the difficulties inherent in the chi-square measure due to its explicit dependence on sample size, Bentler (1990) developed the CFI, which compares the declared model with the worst possible one, and its values estimate the improvement of the declared model. Kline (1998) states that values of .90 and above indicate an adequate fit, while values .95 and greater are indications of an optimum fit (i.e., overlaps with the best possible model by at least 95 percent). Kline (1998) also states that the RMSEA (its purpose is to estimate how much per-parameter error the model contains) values of <.08 suggest adequate fit, and <.05, good fit (overlaps with the worst possible model by at most 5 percent). Both of these indices do not vary due to sample size.

All of the conventional global fit indices indicated an exceptionally well-fitting model: $\chi 2\ (72) = 75.23$, $p = .37$; RMSEA = .01(.90 CI = .000 - .029); and CFI = .995. The chi-square was not significant, the CFI overlapped with the best possible model by 99.5 percent, and the RMSEA indicated that the model overlapped with the worst possible model by less than 1 percent. Almost no inconsistencies among the indices were evident within the model-data fit.

The three types of variables described in the structural equation model and confirmatory factor analysis in Figure 3 (below) are latent variables, observed variables, and error terms. Latent variables (symbolized by ovals) correspond to particular constructs, which in our case are objectives. Observed variables (squares) are actual measurements made

by the quiz questions. The error terms (circles) consist of the measurement error associated with each test item. Kline (1998) does note that standardized factor loadings are far easier to interpret than unstandardized coefficients, since the inter-factor relationships are represented as correlation coefficients rather than covariant estimates of the population.

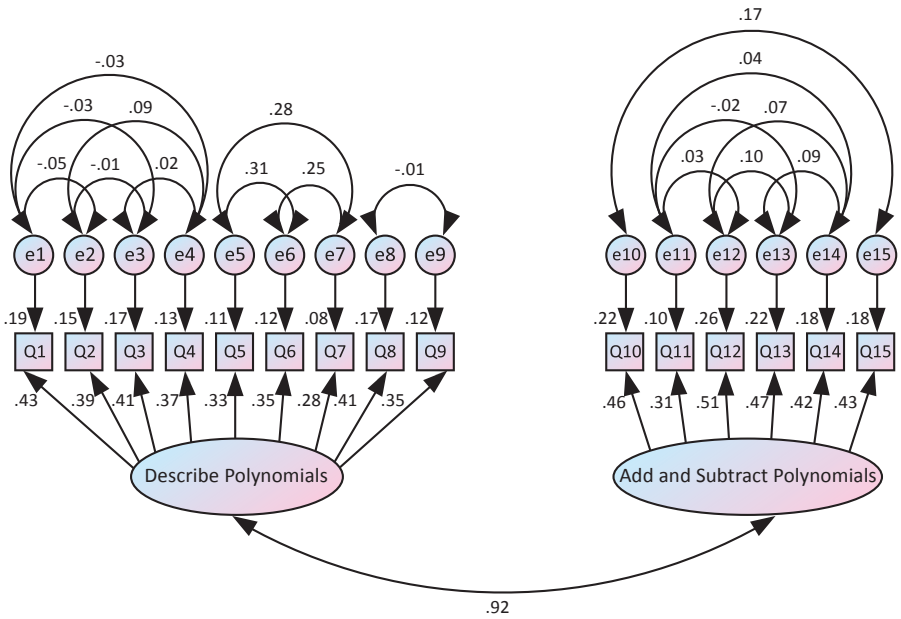

Figure 3. Structural equation model of polynomial quiz.

The relationships between latent variables and their hypothesized observables are characterized by numerical values just to the left of each direct arrow. These values are typically referred to as factor loadings, direct effects, or parameter estimates of path coefficients. They indicate how a change in a latent variable will affect a change in that particular observed variable. Direct effects indicate the expected increase in Y in standard deviation units, controlling for the other predictors. Standardized values should range from 0 to 1, where if a coefficient is .30, it is expected that Y will increase by .30 for every increase of one standard deviation unit in X. Kline (1998) advocates checking estimates for flaws, as they are often indicated by values greater than 1, negative, or non-significant factor loadings. Further, for direct effects that are significant, corresponding standardized values below .10 are considered nil, those between .10 and .30 small, greater than .30 medium, and greater than .50 large (Kline 1998). The range of values for the polynomial quiz was from .28 to .51, indicating medium-to-large direct effects, and the average was .40 (*SD* = .06).

The ovals represent component factor correlations between two latent variables or aligned objectives (also known as indirect effects). In our example, content writers suggested that both topics interact; therefore the correlation between both latent variables will be allowed to vary. Kline (1998) states that standardized component factor correlations for healthy models range from 0 to 1, and that any values that are greater than .80 should be considered substantial, indicating redundancy among observed variables (test questions). Specifically with the polynomial quiz, the correlation between the two objectives was .92, suggesting a great overlap between topics.

The "e#" (e1, e2, e3, etc. within circles) represents the measurement error for each observable (test item) and is the proportion of explained variance. These values are illustrated by values in the upper-left corners of the observed variables, ranging from 0 to 1, where a variance of .20 will indicate that 80 percent of the variance associated with a test item is due to measurement error. As with direct effects, any variance estimates that demonstrate a non-significant effect may be indicative of a structural flaw, and the researcher should consider leaving the variable out of further analysis (or remove it all together) to increase the overall fit of the empirical model (Mueller, 1996). All of the error variances were significant, thus all questions were retained for further analysis.

In post hoc model fitting, many of the error terms may be allowed to correlate with one another to improve model fit if there is a substantive theoretical reason to do so (Kline, 2002). In our case, respective error terms were allowed to correlate, since some test questions are essentially identical. Correlation estimates between error terms e10 and e15 (r = .17), as well as e5 and e6 (r = .31), e6 and e7 (r = .25), and e5 and e7 (r = .28) were the only significant error terms, suggesting overlap between these items.

Edgenuity assessments undergo a rigorous analysis of their reliability and validity.

Discussion and Conclusion

Statistical analyses were conducted to determine the reliability and validity of the polynomial quiz as well as its usefulness as a measure for determining the polynomial conceptual knowledge of middle and high school students. Overall, the initial psychometrics of the polynomial quiz provided positive results. Below is a discussion regarding each general research question. Results are paired with critiques, followed by an examination of the implications.

The polynomial quiz scores obtained a Cronbach's alpha (α) of .75, indicating extensive reliability for the assessment. This result implies there is psychometric support for the polynomial quiz as a reliable tool for middle and high schools.

Results of testing indicated that item indices of congruency ranged from 1.00 to .76, and that all 15 items corresponded to their intended objective. The average congruence index for the quiz was .86 (SD = .10), demonstrating strong content validity. Based on feedback from the judges, the polynomial quiz does have content validity related to measuring conceptual understanding of middle and high school students, and it is a measure that can be easily read and understood by the majority of the intended population.

All of the conventional global fit indices indicate an exceptionally well-fitting model: $\chi 2$ (72) = 75.23, p = .37; RMSEA = .01(.90 CI = .000 - .029); and CFI = .995, suggesting strong construct validity for the quiz. This indicates that the two scales could be used to target specific interventions using Edgenuity content, depending on where a student shows deficiencies.

Results related to the initial psychometrics of the polynomial quiz are extremely promising and indicate it is a sound instrument to measure students' conceptual knowledge within this area of the Algebra I course content.

References

American Educational Research Association (1985). *Standards for educational and psychological testing.* Washington, D.C.: American Psychological Association.

Bentler, P. M. (1990). Comparative fit indexes in structural models. *Psychological Bulletin*, *107*, 238-246.

Clark, L. A., & Watson, D. (1999). Constructing validity: Basic issues in objective scale development. *Psychological Assessment*, *7(3)*, 309-319.

DeVellis, R. F. (2003). *Scale development: Theory and applications.* Newbury Park, CA: Sage.

Haynes, S., Richard, D., & Kubany, E. (1995). Content validity in psychological assessment: A functional approach to concepts and methods. *Psychological Assessment*, *7(3)*, 238-247.

Holden, R. R., Fekken, G. C., & Cotton, D. H. G. (1991). Assessing psychopathology using structured test-item response latencies. *Psychological Assessment, 3,* 111-118.

Kline, R. (1998). *Principles and practice of structural equation modeling.* New York: Guilford Press.

Kuder, G., & Richardson, M. (1937). The theory of the estimation of test reliability. *Psychometrika*, *2(3)*, 151-160.

Lei, P., & Wu, Q. (2007). An NCME instructional module on introduction to structural equation modeling: Issues and practical considerations. *Educational Measurement: Issues and Practice*, *26*, 33-43.

Messick, S. (1993). Validity. In R. L. Linn (Ed.). *Educational measurement* (pp.105-146). Phoenix, AZ: Oryx Press.

Mueller, R. (1996). *Basic principles of structural equation modeling: An introduction to Lisrel and Eqs.* New York: Springer-Verlag Publications , p.82.

Robinson J. P., Shaver P.R. & Wrightsman L. S. (1991). *Measures of personality and social psychological attitudes.* San Diego, CA: Academic Press.

Rovinelli, R., & Hambleton, R. (1977). On the use of content specialists in the assessment of criterion-referenced test item validity. *Dutch Journal of Education Research, 2*, 49-60.

Schumacker, R. E., & Lomax, R. G. (2010). *A beginner's guide to structural equation modeling* (3rd ed.). New York: Routledge.

Thorndike, R. (1997). *Measurement and evaluation in psychology and education.* Upper Saddle River, NJ: Prentice Hall.