

An exploratory study of the statistical and educational implications of
violations of the assumptions of parametric analysis techniques

David A. Wiley, II

C. Victor Bunderson

Brigham Young University

The Edumetrics Institute

Joseph A. Olsen

Brigham Young University

Introduction

Statistical procedures come with assumptions that must be met before they can be used to make valid inferences or test hypotheses. One of the assumptions that must be met for the valid use of parametric statistics can be traced back to Stevens' (1946) seminal article on measurement scaling theory and its taxonomy of measurement scales: nominal, ordinal, interval, and ratio. Parametric procedures assume that the data to be operated upon are interval or ratio, as opposed to being nominal or ordinal (Daniel, 1990). This point was elaborated by Siegel (1956):

In the computation of parametric tests, we add, divide, and multiply the scores from the samples. When these arithmetic processes are used on scores which are not truly numerical, they naturally introduce distortions in those data and this

throw in doubt any conclusions from the test. Thus it is permissible to use the parametric techniques only with scores that are truly numerical.

The implications of this claim are significant for social scientists, who normally gather ordinal and/or categorical data and yet persist in using parametric procedures. This supposed danger of using non-interval data with parametric statistics has been discussed frequently in the literature, with positions ranging from the desperate need for interval measures (Marcus-Roberts & Roberts, 1987; Wright, 1999; Smith, 1999) to the dismissal of the notion that a significant problem exists (Borgatta & Bohrnstedt, 1980; Baker, Hardyck, & Petrinovich, 1966). The purpose of this study is to explore the difference between the parametric analysis of interval and non-interval forms of the same data and the implications of those differences for education.

In order to answer the question regarding the possibly different results from the use of interval measures as opposed to non-interval data, a suitable extant data set was identified. These data were originally gathered by Bunderson (1965) as part of a concept learning study. The problem format is briefly summarized below. More detailed descriptions are available elsewhere (Bunderson, 1965; Bunderson, 1967).

The problems used in the Bunderson (1965) study were adapted from those used originally by Glanzer, Huttenlocher, and Clark (1963), which they referred to as "complex positive" and "complex negative." Eighteen problems were presented in 6 groups of three, each group containing two complex negative instances and a single complex positive instance.

The general format of the complex negative and complex positive problems was as follows: by means of a slide displayed on a screen in a semi-darkened examination room, students were presented eight sets of eight geometric figures either black or white in color. An example eight sets of eight figures (which make one instance or problem) is represented in Figure 1, taken from Bunderson (1967, p. 14). Students were instructed that a "secret combination" of four figures, each in a specific color, existed. The secret combination was present in the first set of figures shown in the series, along with four other irrelevant figures. The students were then shown each set of figures in the series one at a time, along with the word "yes" if all four figures were present in the correct color, or "no" if they were not. Only one bit of information (the color of one of the figures) was changed from the secret combination from set to set within a series. After viewing eight such sets of figures, the students were finally asked to identify the "secret combination." Specially developed answer pads were distributed to students for use in solving the problems. An example is included as Figure 2.

| Number | EXAMPLES | | | | | | | Type | |
|--------|----------|--|--|--|--|--|--|------|-----|
| 1 | | | | | | | | | yes |
| 2 | | | | | | | | | yes |
| 3 | | | | | | | | | no |
| 4 | | | | | | | | | no |
| 5 | | | | | | | | | yes |
| 6 | | | | | | | | | yes |
| 7 | | | | | | | | | no |
| 8 | | | | | | | | | no |

Figure 1. A sample complex negative problem from the Bunderson study. Subjects were shown each line separately and asked to identify the "secret combination."

Complex positive instances were composed of three "no" sets and five "yes" sets of figures. In other words, for these instances it was possible to identify the secret combination using only positive information. Complex negative instances, on the other hand, were composed of four "yes" and four "no" sets, and could only be solved through the use of negative information.

The data analyzed for the current study began as individual student raw scores on the 18 problems described above. Raw scores are simply the sum of the correct responses where correct items are scored with a 1 and incorrect items are scored with a 0. Using a

recently developed process called "Domain Theory" (Bunderson, Newby, & Wiley, in preparation), the original item scores were transformed into interval measures. The process relies on the model proposed by Rasch (1960) for converting dichotomous and rating scale observations into interval measures. This involves the determination of item difficulty and subsequent anchoring of these difficulties across longitudinal cycles as individual's raw scores for each cycle are transformed into interval measures. This was accomplished using the QUEST software package. QUEST employs the UCON and maximum likelihood algorithms. At this point both six raw scores and six measures existed for each of 145 subjects, one at each of the six longitudinal cycles.










| | | | | | | | | | |
|--|---|---|---|---|---|---|---|---|---|
| Student N ^o 201 | | | | | | | | | |
| PROBLEM 1 | 1 |  |  |  |  |  |  |  |  |
|  4 | 2 | w b n <input type="checkbox"/> | w b n <input type="checkbox"/> | w b n <input type="checkbox"/> | w b n <input type="checkbox"/> | w b n <input type="checkbox"/> | w b n <input type="checkbox"/> | w b n <input type="checkbox"/> | |
| | 3 | w b n <input type="checkbox"/> | w b n <input type="checkbox"/> | w b n <input type="checkbox"/> | w b n <input type="checkbox"/> | w b n <input type="checkbox"/> | w b n <input type="checkbox"/> | w b n <input type="checkbox"/> | |
| | 4 | w b n <input type="checkbox"/> | w b n <input type="checkbox"/> | w b n <input type="checkbox"/> | w b n <input type="checkbox"/> | w b n <input type="checkbox"/> | w b n <input type="checkbox"/> | w b n <input type="checkbox"/> | |
| | 5 | w b n <input type="checkbox"/> | w b n <input type="checkbox"/> | w b n <input type="checkbox"/> | w b n <input type="checkbox"/> | w b n <input type="checkbox"/> | w b n <input type="checkbox"/> | w b n <input type="checkbox"/> | |
| | 6 | w b n <input type="checkbox"/> | w b n <input type="checkbox"/> | w b n <input type="checkbox"/> | w b n <input type="checkbox"/> | w b n <input type="checkbox"/> | w b n <input type="checkbox"/> | w b n <input type="checkbox"/> | |
| | 7 | w b n <input type="checkbox"/> | w b n <input type="checkbox"/> | w b n <input type="checkbox"/> | w b n <input type="checkbox"/> | w b n <input type="checkbox"/> | w b n <input type="checkbox"/> | w b n <input type="checkbox"/> | |
| | 8 | w b n <input type="checkbox"/> | w b n <input type="checkbox"/> | w b n <input type="checkbox"/> | w b n <input type="checkbox"/> | w b n <input type="checkbox"/> | w b n <input type="checkbox"/> | w b n <input type="checkbox"/> | |
| The secret combination is: | | w b <input type="checkbox"/> | w b <input type="checkbox"/> | w b <input type="checkbox"/> | w b <input type="checkbox"/> | w b <input type="checkbox"/> | w b <input type="checkbox"/> | w b <input type="checkbox"/> | |

Figure 2. A sample answer pad. Students were instructed to mark the color information they felt was pertinent and use it in order to identify the "secret combination."

Method

To compare the results of using these same data in the two formats, analyses were performed extending the original work conducted by Bunderson. One of Bunderson's untested hypotheses dealt with learner growth across the six cycles. Latent growth modeling (LGM) was used to examine the shape of growth and the fit of data to that shape. Another of Bunderson's hypotheses dealt with the existence of latent classes in the data. Growth mixture modeling was performed to identify latent classes using Mplus software, and LGMs were then developed for the classes. Finally, the LGMs were used with both raw scores and interval measures to investigate relative model fit.

Latent growth modeling techniques provide methods for testing the relative fit of different models on the same data when one of the models is nested within the other. No formal method exists, however, for testing the difference in fit of alternative scalings of the same data with the same model, as was proposed in this study. The reader is therefore asked to inspect the evidence and arguments laid out before him, and judge for himself regarding this hypothesized difference.

Results and Discussion

Latent Growth Model

First an unspecified LGM was developed for the raw score data (see Figure 3). This model was then estimated using the AMOS software program for both the raw score data and the measure data. The parameter estimates are presented in Table 1, and the measures of fit are reported in Table 2.

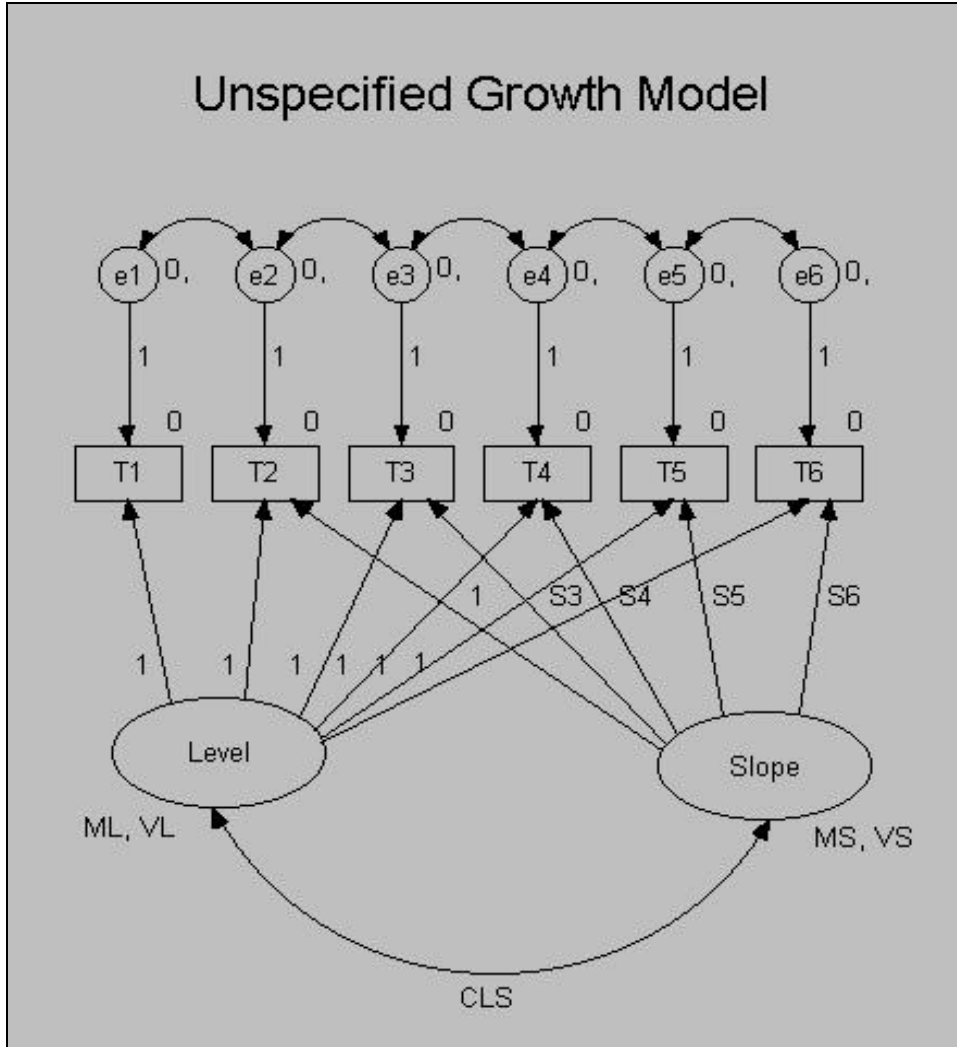


Figure 3. The unspecified growth model for the subjects solving Bunderson's concept problems.

The model presented in Figure 3 represents an attempt to understand the shape of student growth throughout the six cycles. Variables T1 - T6 represent the six time periods or cycles during which students engaged the concept problems. Some error is involved with measuring the subjects at each of these times, and this error is modeled by e1 - e6 (located above the each T variable). The arrows between adjacent error terms indicate the

expectation that standard errors would be correlated, as this is essentially a repeated measures design. At the bottom of the model, the Level represents subject's initial performance, while the Slope variable is a measure of the change of their performance over time. Level and Slope are also correlated since a relationship between initial performance and growth of performance throughout the six cycles was anticipated. The slope estimates are graphed in Figure 4, showing the overall growth curve for all 145 subjects.

Figure 4 shows the growth of subject performance on the concept problems over time, which begins as a rather linear trend that levels off toward the end. Unsurprisingly, this closely resembles the top of the classic S-shaped learning curve.

Table 1

Parameter Estimates for the Unspecified Growth Model

| Parameter | Estimate | Std Err | p-value* |
|-----------|----------|---------|----------|
| ML | -1.198 | 0.177 | 0.000 |
| MS | 1.168 | 0.115 | 0.000 |
| S3 | 1.512 | 0.112 | 0.000 |
| S4 | 2.067 | 0.175 | 0.000 |
| S5 | 2.419 | 0.200 | 0.000 |
| S6 | 2.475 | 0.206 | 0.000 |
| VL | 4.239 | 0.745 | 0.000 |
| VS | 0.660 | 0.178 | 0.000 |
| CLS | 0.374 | 0.279 | 0.180 |

* alpha = .05

Table 2

Latent growth model fit for raw scores and interval measures on the same model

| Fit Measure | Raw Scores | Measures |
|------------------------------------|------------|----------|
| Degrees of Freedom | 7 | 7 |
| Discrepancy | 19.805 | 10.411 |
| RMSEA | 0.113 | 0.058 |
| P for perfect fit | 0.006 | 0.166 |
| P for test of close fit | 0.037 | 0.367 |
| Akaike information criterion (AIC) | 59.805 | 50.411 |

As displayed in Table 2 and the other tables in this report, "discrepancy" is a chi-squared test statistic indicating model fit, with zero indicating perfect fit. The RMSEA is the root mean square error approximation, and zero indicates perfect fit here as well. Browne and Cudeck (1993) have suggested that a value of 0.05 or lower indicates "close" fit, while they "would not want to employ a model with a RMSEA greater than 0.1". P for perfect fit is the p-value of the test that the RMSEA is equal to zero, while P for test of close fit is the p-value for the test that the RMSEA is less than 0.05. Because (a) the null hypothesis is that RMSEA=0, (b) zero indicates perfect fit, and (c) significant p-values would represent evidence that the RMSEA was not zero, in model fitting terms non-significance is desirable from these two statistics. Finally, the AIC is a modification of the goodness of fit chi-squared statistic that incorporates a penalty for model complexity, i.e., it favors simpler models that provide similar fit.

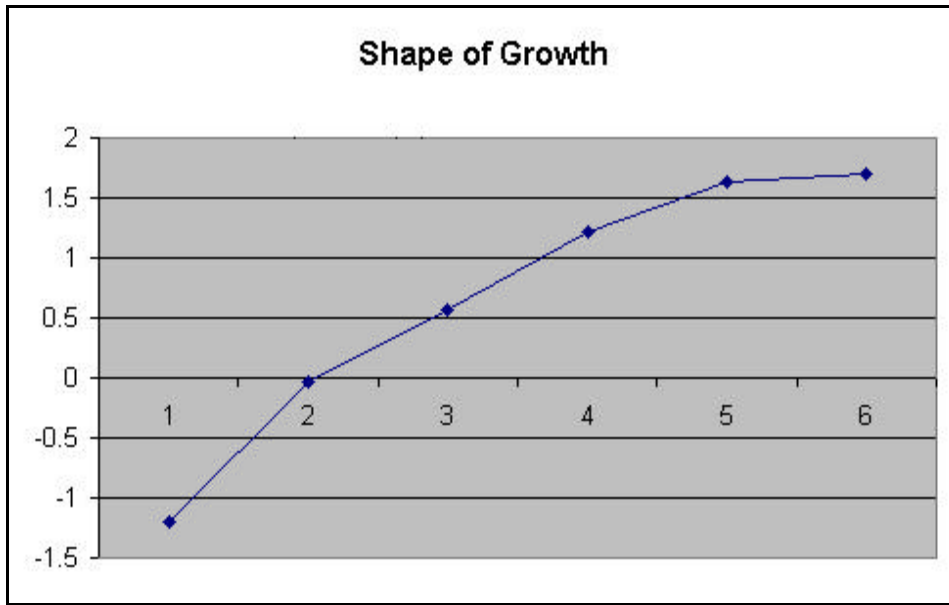


Figure 4. Unit change over time in subject performance on the concept problems plotted using implied means.

The fit indices reported in Table 1 show very clearly that interval measures fit the growth model better than the raw scores. In fact, following the Browne and Cudeck (1993) recommendations for interpreting RMSEA as a fit measure, the model would be rejected when used with raw scores. With measures, on the other hand, the model is near what they described as "close fit." Both the p-values indicate significance at the $\alpha = .05$ level for raw scores (i.e., they show significant odds that the RMSEA does not equal 0), while the p-values for the measures are both insignificant at $\alpha = .05$ (i.e., they do not show significant odds that the RMSEA does not equal 0). Finally, the AIC is also lower for the measures than the raw scores, indicating that in this case the measures not only exhibit superior fit, they would also lead to very different decisions in hypotheses testing scenarios than would the raw scores.

Latent Classes

Because the indicators were continuous, a mixture model was used to identify latent classes in the Bunderson data set. Bunderson had originally hypothesized the existence of three classes within the data. However, when a three-class model was estimated, two classes very similar in shape and separated by a very small distance were discovered. This seemed to suggest the existence of only two classes within the data. A 95% confidence interval error bar graph of the two-class analysis over the six cycles, using interval measures, is presented in Figure 5.

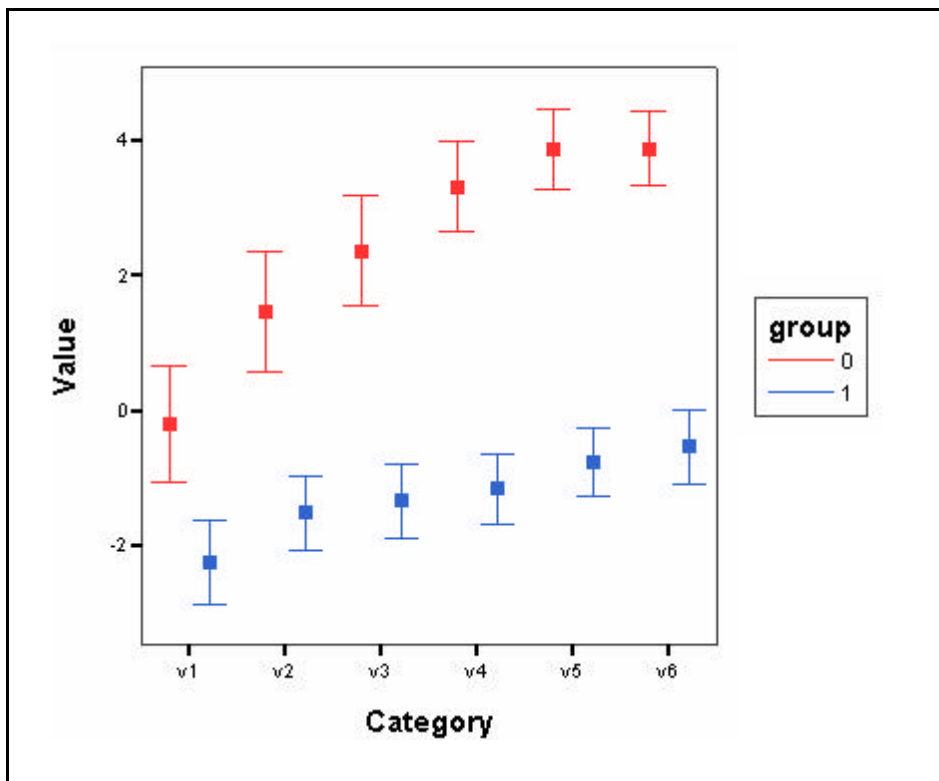


Figure 5. Error bar graph for two latent classes in the Bunderson data.

Figure 5 shows two quite distinct groups within the data. Roughly speaking, some people learn how to solve the problems (the "high class") and others do not (the "low class"). The growth of each of these classes is described by a different growth curve. It was hypothesized that a quadratic LGM would fit the high class's growth, while a linear equation would roughly fit the low class's growth. Group membership probability was recorded for each individual, and the individuals were assigned to one of the two classes using this probability information. Separate LGMs were estimated for the two classes, using both raw scores and interval measures.

The quadratic growth model for the high class is presented in Figure 6, and fit measures for both raw scores and measures are reported in Table 3.

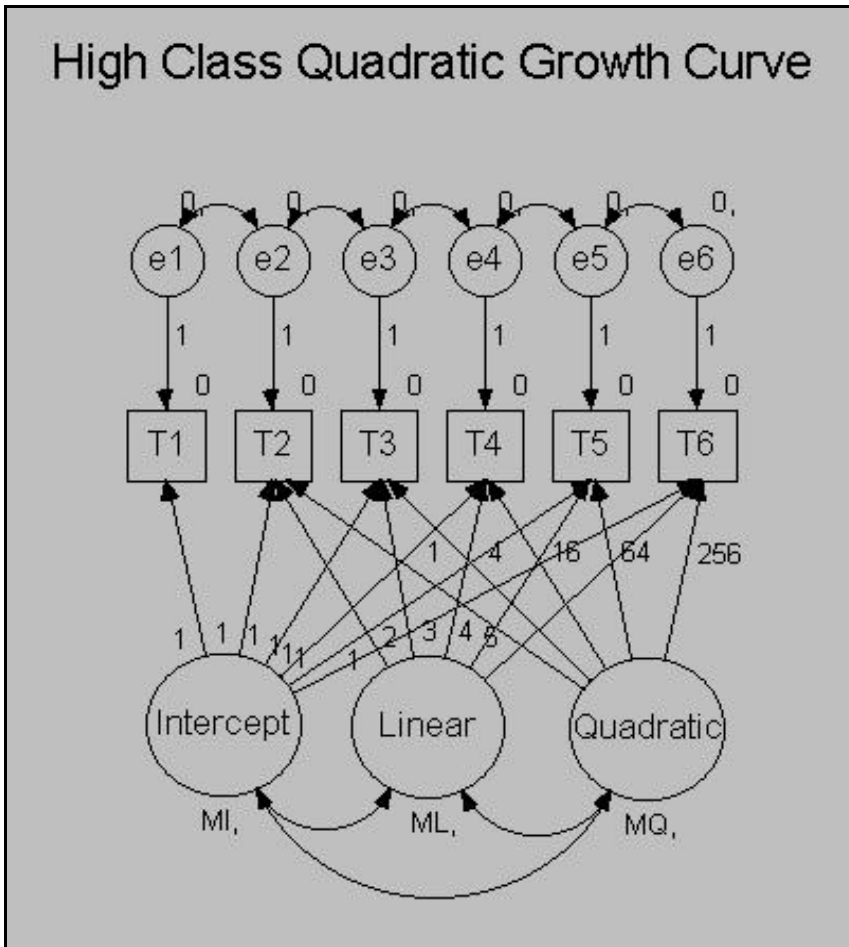


Figure 6. Quadratic growth model for the high class.

The model presented in Figure 6 differs in important ways from the model presented in Figure 3. Instead of simply specifying a "Slope" factor and letting the data determine the shape of the curve, this model fits an equation incorporating both linear and quadratic trend components. The fixed factor loadings from the latent variable Linear to the six time periods (the factor loading on T1 is 0, and thus not shown in the graph) define the linear trend. The fixed factor loadings from the latent variable Quadratic to the six time periods define the quadratic trend. The estimate of the mean for the variable

Quadratic was significant at the $\alpha = .05$ level, and thus its inclusion in the model is needed to fit the data adequately.

Table 3

Comparisons of the fit of the high class raw scores and interval measures on the same

LGM

| Fit Measure | Raw Score | Measures |
|------------------------------------|-----------|----------|
| Degrees of Freedom | 11 | 11 |
| Discrepancy | 27.811 | 15.062 |
| RMSEA | 0.198 | 0.125 |
| P for perfect fit | 0.000 | 0.035 |
| P for test of close fit | 0.001 | 0.078 |
| Akaike information criterion (AIC) | 67.811 | 55.062 |

As above, each of the fit measures based on the fit of interval measures to the model is superior to those based on the fit of raw scores.

Finally, the linear growth model for the low class is presented in Figure 8, and fit measures for both raw scores and measures are reported in Table 3.

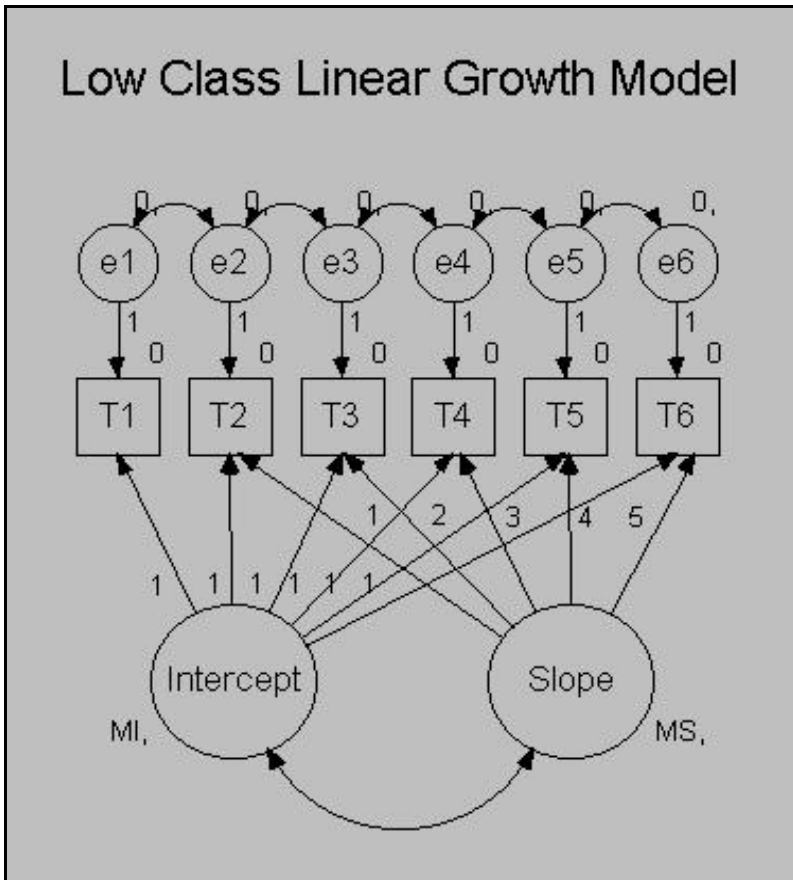


Figure 8. Linear growth model for the low class.

This model is very similar to that presented in Figure 3, except that the Slope is defined as a simple linear trend.

Table 4

Comparisons of the fit of the low class raw scores and interval measures on the same

LGM

| Fit Measure | Raw Score | Measures |
|------------------------------------|-----------|----------|
| Degrees of Freedom | 11 | 11 |
| Discrepancy | 21.300 | 23.012 |
| RMSEA | 0.118 | 0.126 |
| P for perfect fit | 0.030 | 0.018 |
| P for test of close fit | 0.074 | 0.049 |
| Akaike information criterion (AIC) | 53.300 | 55.012 |

In the final case, the difference between scores and measures all but disappears, with the raw scores, if anything, fittingly slightly better than the measures for this class. This result does not agree with theoretical predictions. Further study will be necessary to understand why this change in degree of fit occurs.

Conclusions and Recommendations

Latent growth modeling has application in educational settings for modeling the growth trajectory of individuals over time. This type of information can be invaluable in providing feedback to both the student and instructor regarding manners in which they can improve the effectiveness of their learning and instruction. This information could also be of great use to computer-adaptive testing and instructional systems. As the mixture model analysis for latent classes performed in this study shows, radically different classes of learners can exist within a data set that fits a single class growth model very well. Providing "one-size-fits-all" instructional support based on a single, "averaged" model of growth has the potential to be instructionally hazardous. For

example, high performing learners could become frustrated by too much structure and guidance from the instructor or CAI system. At the same time, learners who are "just not getting it" could become frustrated as the instructor or computer assumes they belong to a class that is performing quite well and growing rapidly in ability, and therefore does not provide adequate support or help. The identification of latent classes and their individual growth trajectories is essential to the instructional well-being of those we serve as educators, and a powerful tool under utilized even by professional instructional designers.

If assessment is to play as large and successful a role in education as instructional design textbooks claim it should, tools and procedures for making accurate measurements must be developed. Even in the social sciences, where non-interval data is utilized almost exclusively, parametric statistics are frequently used as a basis for instructional and other decisions. This paper has provided cases in which the difference between the rejection and retention of the null hypothesis depended on the type of data used (either raw scores or measures), showing that parametric statistics are not as robust to violations of the assumption of interval data as some have previously believed. If it can be assumed that data are gathered and analyzed in order to support instructional decision making, then -- in at least some cases -- the fate of students depends on whether their instructor takes the steps necessary to meet the most fundamental assumptions of the statistics used as a basis for the decisions made.

It is therefore the authors' recommendation that high stakes instructional and assessment systems, particularly computer-based or automated systems, utilize interval measures in their calculations. Bunderson, Newby, and Wiley (in preparation) describe a method of transforming raw scores to interval measures based on the Rasch model, and other methods will certainly be recommended in the literature over time. Because the

calibration and other computational activities involved in construct scaling theory are rather involved, the linear measures are not recommended for non-high stakes scenarios, for example, in the calculation of grades for a high school political science class.

References

- Baker, B. O, Hardyck, C. D., & Petrinovich, L. F. (1966). Weak measurements vs. strong statistics: An empirical critique of S. S. Stevens' proscriptions on statistics. Educational and Psychological Measurement, 26 (2), 291-309.
- Borgatta, E. F. & Bohrnstedt, G. W. (1980). Level of measurement: once over again. Sociological Methods and Research, 9 (2), 147-160.
- Bunderson, C. V. (1965). Transfer functions and learning curves: The use of ability constructs in the study of human learning. Research Bulletin 64-62. Princeton, NJ: Educational Testing Service.
- Bunderson, C. V. (1967). Transfer of mental ability at different stages of practice in the solution of concept problems. Research Bulletin RB-67-20. Princeton, NJ: Educational Testing Service.
- Browne, M., & Cudeck, R. (1993). Alternative ways of assessing model fit. In K. A. Bollen & J. S. Long (Eds.), Testing structural equation models (pp.136-162). Newbury Park, CA: Sage.
- Daniel, W. W. (1990). Applied nonparametric statistics. PWS-KENT Publishing Company, Boston.
- Glanzer, M., Huttenlocher, J. & Clark, W. H. (1963). Systematic operations in solving concept problems: A parametric study of a class of problems. Psychology Monograph, 77.
- Lazarsfeld, P. F., and Henry, N. W. (1968). Latent structure analysis. Boston: Houghton Mifflin.

- Marcus-Roberts, H. M. & Roberts, F. S. (1987). Meaningless statistics. Journal of Educational Measurement, 12 (4), 383-394.
- Rasch, G. (1960). Probabilistic models for some intelligence and attainment tests. Copenhagen, Denmark: Danmarks Paedagogiske Institute. (reprinted, Chicago, IL: University of Chicago Press, 1980)
- Siegel, S. (1956). Nonparametric methods for the behavioral sciences. New York: McGraw-Hill.
- Smith, R. M. (1999). Rasch measurement models. Chicago: MESA Press.
- Stevens, S. S. (1946). On the theory of measurement scales. Science, 103, 677-680.
- Wright, B. D. (1999). Fundamental measurement for psychology, in S. E Embretson & S. L. Hershberger, (Eds.). The new rules of measurement: What every psychologist and educator should know (pp. 65-104). Mahwah, NJ: Lawrence Erlbaum.
- Wright, B. D., and Douglas, G. A. (1996). Estimating measures with known item difficulties. Rasch Measurement Transactions [On-line.] Available: <http://www.rasch.org/rmt/rmt102.htm#Estimating>