Multiple Linear Regression Viewpoints

A Publication sponsored by the American Educational Research Association's Special Interest Group on Multiple Linear Regression: The General Linear Model

MLRV

Volume 36 • Number 2 • 2011

Table of Contents

Multiple Linear Regression Viewpoints

Multiple Linear Regression Viewpoints (*MLRV*) is a publication sponsored by the American Educational Research Association's Special Interest Group on Multiple Linear Regression: The General Linear Model (SIG/MLR: GLM). It is published twice a year to facilitate communication among professionals who focus their research on the theory, application, or teaching of multiple linear regression models and/or the general linear model. Manuscripts submitted to *MLRV* should conform to the language, style, and format of the *Publication Manual of the American Psychological Association* (6th ed., 2010). Manuscripts should be prepared in Word, be doubled-spaced, use 12 font, contain a 100 word abstract, have author(s) identifying information appear on the title page only, and consist of no more than 30 pages in length (including equations, footnotes, quotes, and references). Mathematical and Greek symbols should be clear and concise. Tables, figures, and diagrams must be photo copy ready for publication. All manuscripts should be submitted electronically to the editor.

Once received by the editor, manuscripts will be anonymously peer-reviewed by two editorial board members. The review process will take approximately 2 to 3 months. A letter acknowledging receipt of the manuscript will be sent to the first author, and upon review completion, a letter indicating the peer-review decision will be sent to the first author. Potential authors are encouraged to contact the editor to discuss ideas for contributions or determine if their manuscript is suitable for publication in *MLRV*.

EDITOR

David Walker
Northern Illinois University
College of Education
318 Graham Hall
DeKalb, IL 60115
Phone: (815) 753-9362
Fax: (815) 753-2100
Email: dawalker@niu.edu

ASSOCIATE EDITOR

T. Mark Beasley
University of Alabama-Birmingham
Department of Biostatistics
School of Public Health
Ryals Public Health Bldg.
Birmingham, AL 35294
Phone: (205) 975-4957
Fax: (205) 975-2540
Email: MBeasley@ms.soph.uab.edu

ORDER INFORMATION

Cynthia Campbell, Managing Editor
Northern Illinois University
Department of Educational Technology, Research and Assessment
DeKalb, IL 60115
Phone: (815) 753-8471
Fax: (815) 753-9388
Email: ccampbell@niu.edu

Multiple Linear Regression Viewpoints

David A. Walker, Editor
Northern Illinois University

T. Mark Beasley, Associate Editor
University of Alabama-Birmingham

Isadore Newman, Editor Emeritus
Florida International University

Randall E. Schumacker, Editor Emeritus
University of Alabama-Tuscaloosa

Editorial Board

Multiple Linear Regression Viewpoints (ISSN 0195-7171) is published by the AERA Special Interest Group on Multiple Linear Regression: General Linear Model through **Northern Illinois University** and the **University of Alabama-Birmingham**.

Subscription and SIG membership information can be obtained from:
Cynthia Campbell, Managing Editor
Department of Educational Technology, Research & Assessment
Northern Illinois University
DeKalb, IL 60115-2854
ccampbell@niu.edu

MLRV abstracts appear in CIJE, the ERIC system, and microform copies are available from University Microfilms International, 300 North Zeeb Road, Ann Arbor, MI 48106. *MLRV* is listed in the *EBSCO Librarians Handbook*.

Demonstration of How Score Reliability is Integrated Into SEM and How Reliability Affects All Statistical Analyses

Z. Ebrar Yetkiner
Texas A & M University

Bruce Thompson
Texas A & M University – Baylor College of Medicine

SEM is a generalization of multiple regression within the General Linear Model (GLM). The purpose of the present paper is to explain and illustrate (a) how score reliability estimation is integrated into structural equation modeling (SEM) and (b) how score reliability affects all statistical analyses within the GLM. A heuristic SEM model using the Holzinger and Swineford (1939) data is utilized for these purposes. This introductory tutorial includes the necessary data files and software syntaxes so that the interested reader can readily replicate and further explore these important foundational statistics concepts.

Structural equation modeling (SEM) is a marriage of the multiple regression path analytic modeling promulgated by Sewell Wright (1921, 1934) in the 1920s with quantitative methods that estimate the psychometric reliabilities of scores. SEM allows for the evaluation of rival causal path directions and both direct and indirect effects. As such, SEM is an extension of multiple regression as the General Linear Model (see Cohen, 1968; Zientek & Thompson, 2009).

Quantitative researchers frequently speak of "error," and the importance of minimizing "error" within studies. However, there are three distinct types of error that are potentially at play within statistical analyses. For purposes of clarity perhaps the generic term "error" should be abandoned in favor of the more precise clear specification of exactly which error type is being considered in any given instance.

First, *sampling error* is any feature of sampled data that misrepresent in any way the dynamics in the population from which the sampled data were drawn (Thompson, 2006a). Every sample necessarily is "fluky" to some degree. Only the population has no flukiness in representing itself. As the statistical cliché states, "Samples are like people, and every single one is individual or idiosyncratic, and some are very idiosyncratic!" We can see sampling error at work if we draw repeated samples of a given size from a population, compute whatever statistic interests us (e.g., the mean, the median, the Huber estimate, the Pearson *r*), and then plot these estimates as a *sampling distribution*.

And we can compute the standard deviation of the sampling distribution (i.e., the *standard error of the statistic*, such as the SE_M or the SE_r) to quantify how much idiosyncracy we expect, on average, for a given statistic computed for samples of a given sample size. Sampling errors for all statistics tend to get larger as sample sizes get smaller.

Sampling error tends to inflate the magnitudes of many of the effect sizes (e.g., r^2, R^2, η^2; see Thompson, 2006b, 2007) that today are the primary vehicles for interpreting quantitative research results (Wilkinson & American Psychological Association [APA] Task Force on Statistical Inference, 1999). However, effect size estimates can be corrected for sampling error influences by subtracting out of the initial unadjusted estimates the estimated amounts of sampling error (e.g., adjusted $\underline{r}^2$, adjusted $\underline{R}^2$, ω^2; see Wang & Thompson, 2007). Thus, quantifying the likely amount of sampling error becomes important to the endeavor of correcting effect size estimates so that our effect size estimates are more accurate.

Second, *model specification error* "occurs when (a) predictor variables that should not be used are included in the model, or (b) necessary predictor variables are omitted, or (c) the incorrect analysis is used (e.g., a linear form of relationships is modeled when relationships are curvilinear or logistic)" (Thompson, 2006a, p. 247). Every statistical analysis tests the fit of data to a model (Zientek & Thompson, 2009), and indeed effect sizes can be conceptualized as quantifying fit (Thompson, 2006a). Thus, in a sense, the Mean $\text{Square}_{\text{ERROR}}$ or 1 - some effect sizes (e.g., 1 - either r^2 or ω^2 or the SEM NFI or CFI fit statistic) can be conceptualized as quantifying the degree of model misspecification in an analysis.

Of course, when we interpret these model misspecification estimates, we must remember that we do not expect most models to be perfectly specified. In fact, we typically develop theory to explain certain phenomena, but we want theory to be simple. Because theories are developed to be parsimonious (i.e., to oversimplify reality, but still be useful), we usually expect theory to be imperfect or to some degree untrue. Thus, we usually do not expect actual data to perfectly fit models.

Third, *measurement error* is the error arising in scores because all scores inherently are imperfectly measured. As explained by Thompson (2003b),

> [M]any of us begin our day by stepping on a scale to measure our weight. Some days when you step on your bathroom scale you may not be happy with the resulting score. On some of these occasions, you may decide to step off the scale and immediately step back on to obtain another estimate. If the second score is half a pound lighter, you may irrationally feel somewhat happier, or if the second score is slightly higher than the first, you may feel somewhat less happy.
>
> But if your second weight measurement yields a score 25 pounds lighter than the initial measurement, rather than feeling happy, you may instead feel puzzled or perplexed. If you then measure your weight a third time, and the resulting score is 40 pounds heavier, you probably will question the integrity of all the scores produced by your scale. It has begun to appear that your scale is exclusively producing randomly fluctuating scores.
>
> In essence, your scale measures "nothing." That is, measurement protocols measure "nothing" when the scores they produce are completely unrelated to any and all systematic or nonrandom dynamics...
>
> When measurements yield scores measuring "nothing," the scores are said to be "unreliable." At the other extreme, if the measurement yielded scores with no elements whatsoever of random fluctuation, the scores would be described as perfectly reliable. The idea of evaluating the psychometric integrity of scores as regards these influences is not new. Spearman (1904, 1910) articulated the original conceptualizations of "reliability." (p. 4)

Measurement error damages the results from quantitative studies. Measurement error attenuates both the magnitudes of effect sizes and power against Type II errors. For example, Shaw, Kopriva, and Tracz (1993) found that statistical power against Type II error drops precipitously with lower score reliability. This is why computing score reliability in a quantitative study before any other statistical analyses are conducted is so important when using classical statistical analyses (e.g., t test, Pearson r, ANOVA, multiple regression, MANOVA, descriptive discriminant analysis, canonical correlation analysis).

As the APA Task Force on Statistical Inference emphasized so strongly,

> It is important to remember that a test is not reliable or unreliable. Reliability is a property of the scores on a test for a particular population of examinees. Thus, authors should provide reliability coefficients of the scores for the data being analyzed even when the focus of their research is not psychometric. Interpreting the size of observed effects requires an assessment of the reliability of the scores. (Wilkinson & APA Task Force on Statistical Inference, 1999, p. 596)

Thompson and Vacha-Haase (2000), in their article "Psychometrics is datametrics: The test is not reliable," argued strongly that tests are not themselves reliable, and instead reliability is a property of scores, and score reliability for a given measure may fluctuate over samples or time. This is exactly why Vacha-Haase (1998) invented Reliability Generalization (RG) as a measurement meta-analytic method to explore (a) typical reliability of scores from a given measure over administrations, (b) the homogeneity or heterogeneity of score reliabilities for a measure over administrations, and (c) the design or sample features that predict variation in score reliabilities for a measure over administrations (see Henson & Thompson, 2002).

However, even today too few researchers realize that reliability is not a property of tests themselves, and then presume that the reliability reported for scores described in test manuals are the reliabilities for any subsequent administrations of a given measure. As Thompson (2006a) noted, "Unfortunately, too many researchers incorrectly presume that tests are reliable, and (therefore) rarely check the reliabilities of the scores actually being analyzed" (p. 356).

Because poor score reliability inherently compromises quantitative study results, every quantitative report ought to include an analysis of the score reliabilities of the data actually being analyzed in the study. Sadly, decades ago in his empirical study of quantitative research reporting practices, Willson (1980) noted that only 37% of the articles he studied explicitly provided reliability coefficients for the data being analyzed, and he concluded "that reliability... is unreported in... [so much published research] is... inexcusable at this late date" (p. 9).

Two decades later, more contemporary reporting situation remains unsatisfactory. In their *measurement mega-metaanalysis* of prior RG studies, Vacha-Haase, Henson, and Caruso (2002) reported that, of the thousands of journal articles reviewed in prior RG studies, only 24.4% of these thousands of articles cited reliability coefficients for the data actually being analyzed, and a stunning 56.2% of the thousands of journal articles reviewed contained no mention whatsoever of reliability!

General Linear Model (GLM)

The General Linear Model (GLM) is the concept that "all analytic methods are correlational ... and yield variance-accounted-for effect sizes analogous to r^2 (e.g., R^2, η^2, ω^2)" (Thompson, 2000, p. 263). As Graham (2008) explained,

> The vast majority of parametric statistical procedures in common use are part of [a single analytic family called] the General Linear Model (GLM), including the t test, analysis of variance (ANOVA), multiple regression, descriptive discriminant analysis (DDA), multivariate analysis of variance (MANOVA), canonical correlation analysis (CCA), and structural equation modeling (SEM). Moreover, these procedures are *hierarchical* [italics added], in that some procedures are special cases of others. (p. 485)

In 1968, Jacob Cohen explained that multiple regression analysis subsumes all univariate parametric statistical analyses (e.g., t tests, ANOVA, ANCOVA, Pearson r) as special cases. In 1978, Knapp showed that all commonly utilized univariate and multivariate analyses (e.g., Hotelling's T^2, MANOVA, MANCOVA, descriptive discriminant analysis) are special cases of canonical correlation analysis (CCA). Thompson (1984, 1991) and Zientek and Thompson (2009) provide more detail on CCA and the GLM.

These traditional univariate and multivariate parametric analyses dominated quantitative research reports for many years (Willson, 1980), and remain common (Kieffer, Reese & Thompson, 2001). But none of these analyses (e.g., t tests, ANOVA, ANCOVA, Pearson r, Hotelling's T^2, MANOVA, MANCOVA, descriptive discriminant analysis) incorporate estimates of or adjustments for measurement error within the scores being analyzed. In effect, traditional statistical analyses presume perfect or near-perfect score reliability.

More recently, a modern statistical analysis today called structural equation modeling (SEM; see Thompson, 2000, for an accessible introduction) has been developed and popularized in various software packages, such as AMOS (Arbuckle, 2007). In 1981, Bagozzi, Fornell, and Larcker demonstrated that structural equation modeling (SEM) is an even more general case of the GLM (see Fan, 1997 for a cogent explanation).

Even though SEM is the broadest case of the GLM, and subsumes CCA and classical parametric analyses as special cases, unlike CCA and multiple regression and the other classical analyses, SEM does estimate and make adjustments for score reliability estimates as part of the statistical analysis. In other words, SEM is *both* simultaneously a statistical *and* a measurement analysis.

It is important to note that there are various sources of measurement error (e.g., multiple raters, administrations, or test versions) that can result in unreliable scores. Cronbach, Gleser, Nanda, and Rajaratnam (1972) developed Generalizability Theory, which analyzes various measurement errors simultaneously, as a way to quantify each of the numerous sources of measurement error, and their interactions with each other (Thompson, 2003a). SEM is powerful also in that it can account for multiple sources of measurement error that can be modeled within the analysis (DeShon, 1998, p. 412). However, even some experienced SEM researchers may not fully grasp exactly where and how score reliability is estimated within SEM.

Purposes of the Present Paper

The present paper was written as a tutorial to achieve two purposes. First, the paper explicates exactly where and how score reliability is estimated within SEM. Second, the paper illustrates exactly how changes in score reliability estimates can impact all the estimates in a given quantitative analysis within the GLM.

We use actual data, and provide the relevant AMOS computer program syntax files, to enable the interested reader to replicate our results, or to explore the impacts of using even more alternative

reliability estimates on the parameter estimates within the model. We encourage readers to make use of these syntax files, so that understanding of how reliability affects results with the GLM will be real and concrete.

We do not make any claim that statisticians are unaware that measurement error is estimated as part of structural equation modeling. However, applied researchers may not fully understand exactly where and how reliability is estimated as part of SEM, or how reliability may impact parameter estimates with the GLM. The present tutorial *may prove useful in introducing students to these important foundational concepts*.

Illustrative Data

Real data are often used to provide concrete examples in books and journal articles. One dataset widely used in the social sciences in heuristic examples are the iris flower data reported and used by Sir Ronald Fisher (1936) in his explication of discriminant analysis.

A second dataset (Holzinger & Swineford, 1939, pp. 81-91) has been even more widely used in journal articles and textbooks as heuristic data and involves scores of 301 children from two Chicago schools on several dozen tests. These data are widely available on the internet. For example, the data can be retrieved at: http://people.tamu.edu/~bthompson/datasets or http://www.psych.yorku.ca/friendly/lab/files/psy6140/data/psych24r.sas

Here, scores on six tests from the Holzinger and Swineford (1939) data were used:

Variable Name	SPSS Variable Label
T1	VISUAL PERCEPTION TEST FROM SPEARMAN VPT, PART III
T2	CUBES, SIMPLIFICATION OF BRIGHAM'S SPATIAL RELATIONS TEST
T3	PAPER SHAPES THAT CAN BE COMBINED TO FORM A TARGET
T6	PARAGRAPH COMPREHENSION TEST
T7	SENTENCE COMPLETION TEST
T9	WORD MEANING TEST

The first three variables were modeled as measuring Spatial Ability, and the last three variables were modeled as measuring Verbal Ability. The interested reader can replicate the analyses reported here using the data reported in Appendix A and the four AMOS syntax files reported in Appendices B through E.

Results

We conducted four analyses of the six variables using a single model, but with different estimates of score reliabilities within the models (i.e., near perfect score reliability, near zero score reliability, score reliability estimated using the actual data, and score reliability estimated using previously published reliability coefficients). Our simple model focused on estimating the correlation between the latent Spatial Ability and Verbal Ability variables. For example, Figure 1 presents the Input Model used in AMOS to declare what model we were testing and what parameters we wanted estimated in the analysis.

Figure 1 declared that scores on the underlying Spatial Ability latent variable were one cause (i.e., the one-headed arrows go from Spatial Ability to T1, T2, and T3) of the scores observed on the variables T1, T2, and T3, and that scores on the underlying Verbal Ability latent variable were one cause of the scores observed on the variables T6, T7, and T9. Drawing the two-headed arrow between Spatial Ability and Verbal Ability declares that we are modeling that these two latent variables are correlated, and that we want AMOS to estimate this Pearson *r*.

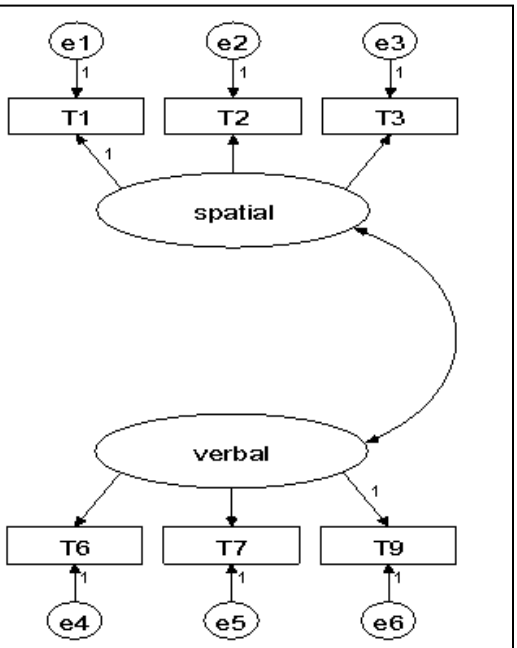

Figure 1. AMOS Input Model with Error Variances Estimated from the Data

The Figure 1 Input Model also declares that the factor pattern coefficient (Thompson, 2004) between T1 and Spatial Ability is 1.0, because the one-headed arrow between T1 and Spatial Ability has a 1.0 typed on the Input Model, and the factor pattern coefficient between T9 and Verbal Ability also is 1.0. This is one way out of many to "identify" the model (see Thompson, 2004), which means to set a scale or metric on the latent construct so that each factor pattern coefficient has only one mathematically correct value.

Think of wanting to measure Height of three people (i.e., Deborah, Wendy, or Carol). We can measure Height in infinitely many metrics. But we must have some (any) metric. We can measure in inches, feet, centimeters, meters, and so forth. We can even invent a completely new metric, as long as our new ruler remains at least intervally-scaled. Or, we can measure in the metric of the people. For example, we can say that we want to measure Height in multiples of 1 Deborah, or half a Deborah. Then, if Wendy is 0.8 as tall as Deborah, Wendy's Height would be 0.8 Deborahs (or 1.6 Deborahs if we are measuring in the metric of half a Deborah). Similarly, if Carol is 1.1 times as tall as Deborah, Carol's height would 1.1 Deborahs (or 2.2 Deborahs if we are measuring in the metric of half a Deborah). Thus, the Figure 1 Input Model declares that we are measuring Spatial Ability in the metric of T1, and we are measuring Verbal Ability in the metric of T9.

The Input Model also declares that each of the six measured variables is measured without perfect reliability, because scores on each measured variable in addition to being caused by a latent variable (either Spatial Ability or Verbal Ability), are also caused by an error variance (i.e., the Figure 1 circles whose names all begin with "e", such as "e1"). Six error variances (i.e., one for each of the six measured variables) are estimated in the Figure 1 Input Model.

If (i.e., if and only if) a model is correctly specified (i.e., there is no model specification error), the error variances are measurement error variances (i.e., V_{ERROR}). So, if the model is correctly specified, the reliability coefficient (r_{XX}) of the measured variable would be:

$$r_{XX} = V_{RELIABLE} / V_{OBSERVED} = [V_{OBSERVED} - V_{ERROR}] / V_{OBSERVED},$$

where $V_{RELIABLE}$ is the variance of the reliable or true score component of the observed scores, and $V_{OBSERVED}$ is the variance of the observed scores on a given variable.

For example, as reported in Appendix A, the standard deviation (*SD*) of the 301 scores on the measured variable T1 was 7.00. Thus, the observed variance of T1 was $V_{OBSERVED} = 7.00^2 = 49.00$. If the V_{ERROR} of T1 was 0.0. the estimated reliability of the 301 T1 scores would be:

$$[49.00 - 0.0] / 49.00 = 1.00 \text{ or } 100\%.$$

Because (a) we can compute and thus know the $V_{OBSERVED}$ for each measured variable, and (b) the SEM program will output the error variances for models having error terms (as do all six of the measured variables in Figure 1), in effect we are estimating the score reliabilities of our measured variables *when we are estimating the error variances of our measured variables*, even when the reliability calculations are not routinely provided in our outputs!

In AMOS we can also fix the error variances of measured variables to be particular numbers. Of course, logically we cannot fix any error variances to be negative, because variances are always squared values, and thus mathematically cannot be negative. We also cannot logically set the error variance of a measured variable to be larger than $V_{OBSERVED}$, because doing so would assert the impossibility that more than the total observed variance in the scores of a measured variable was due to measurement error.

Figure 2 presents an Input Model in which all six error variances were set (and thus not estimated in the Output Model) equal to 0.01. Thus, in this model, the measured variables were
asserted as having been measured with almost perfect reliability. Of course, the reliability coefficients for the six measured variables were not set to be equal, because even though V_{ERROR}'s were all set to equal 0.01, the $V_{OBSERVED}$'s for the six measured variables were all different. For example, the reliability of T1 was set equal to:

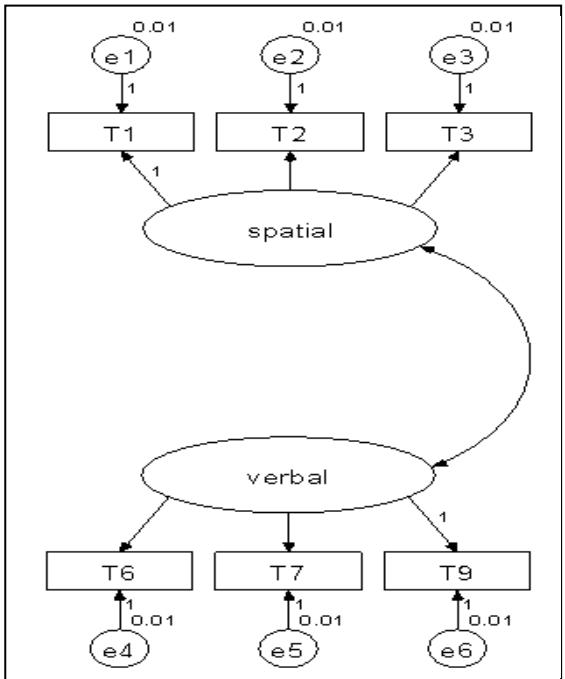

Figure 2. AMOS Input Model with Assumed Near Perfect Score Reliabilities

$$[49.00 - 0.01] / 49.00 = 48.99 / 49.00 = 0.999796,$$

while the reliability of T3 was set equal to:

$$[2.83^2 - 0.01] / 2.83^2 = [8.01 - 0.01] / 8.01 = 8.00 / 8.01 = 0.998752$$

Model #1 of 4: Near Perfect Score Reliability

To illustrate for these data the results when *score reliabilities were nearly perfect*, the six error variances (i.e., the measurement error variances iff the model is presumed to be correctly specified) were all set in the Input Model to be near zero. Specifically, the six error variances for T1, T2, T3, T6, T7 and T9 were all constrained to equal 0.01. Figure 2 is the Input Model for this analysis.

Figure 3 presents the Output Model parameter estimates for the standardized model. The interested reader can reproduce these results by applying the Appendix B AMOS syntax commands to the Appendix A data file. For this analysis, the Output Model estimated correlation between Spatial Ability and Verbal Ability was 0.38. The four unconstrained factor pattern coefficients were 1.00 and 1.00 for T2 and T3 on Spatial Ability, respectively, and 1.00 and 1.00 for T6 and T7 on Verbal Ability, respectively. In this model, declaring the measured variables to have perfect or near perfect score reliability inescapably means that therefore the observed scores on the measured variables can only be due to the underlying Spatial Ability and Verbal Ability constructs, and thus these factor pattern coefficients are all unavoidably estimated to be 1.00's.

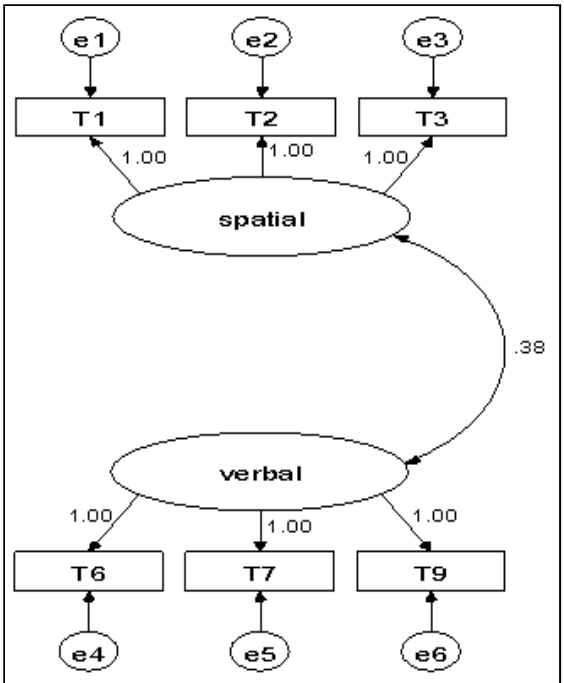

Figure 3. AMOS Output Model with Assumed Near Perfect Score Reliabilities

Model #2 of 4: Zero or Near Zero Score Reliability

To illustrate for these data the results when *score reliabilities were perfectly unreliable*, the six error variances (i.e., the measurement error variances iff the model is presumed to be correctly specified) were all set in the Input Model to equal the observed variances of each measured variable. For example, as reported in Appendix A, the standard deviation (SD) of the 301 scores on the measured variable T1 was 7.004593, and in this analysis the Input Model error variance for T1 was set equal to 7.004593^2, or 49.064319. Similarly, the standard deviation (SD) of the 301 scores on the measured variable T2 was 4.709802, and in this analysis the Input Model error variance of T2 was set equal to 4.709802^2, or 22.182237.

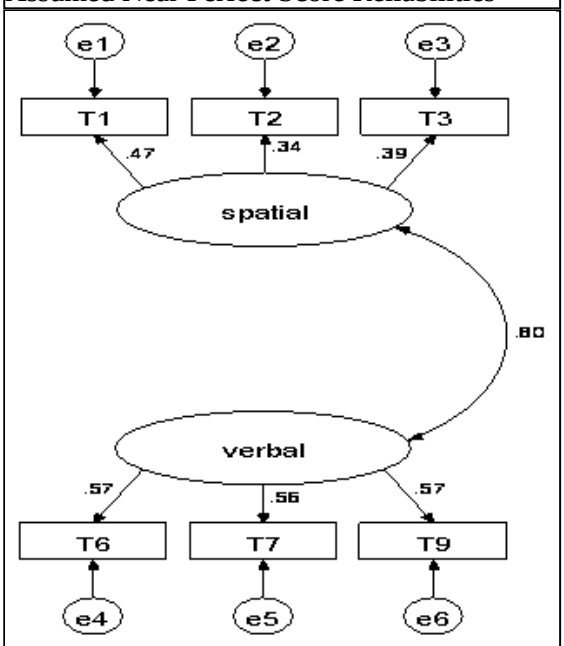

Figure 4. AMOS Output Model with Assumed Near Zero Score Reliabilities.

Figure 4 presents the Output Model parameter estimates for the standardized model. The interested reader can reproduce these results by applying the Appendix C AMOS syntax commands to the Appendix A data file. For this analysis, the Output Model estimated correlation between Spatial Ability and Verbal Ability was 0.80. The four unconstrained factor pattern coefficients were 0.34 and 0.39 for T2 and T3 on Spatial Ability, respectively, and 0.57 and 0.56 for T6 and T7 on Verbal Ability, respectively.

Model #3 of 4: Score Reliability Estimated Within the Analysis

To illustrate for these data the results when *score reliabilities were estimated within the analysis*, the six error variances (i.e., the measurement error variances if the model is presumed to be correctly specified) in the Input Model were all freed to be estimated. Figure 1 is the Input Model for this analysis.

Figure 5 presents the Output Model parameter estimates for the standardized model. The interested reader can reproduce these results by applying the Appendix D AMOS syntax commands to the Appendix A data file. For this analysis, the Output Model estimated correlation between Spatial Ability and Verbal Ability was 0.52. The four unconstrained factor pattern coefficients were 0.41 and 0.50 for T2 and T3 on Spatial Ability, respectively, and 0.85 and 0.85 for T6 and T7 on Verbal Ability, respectively.

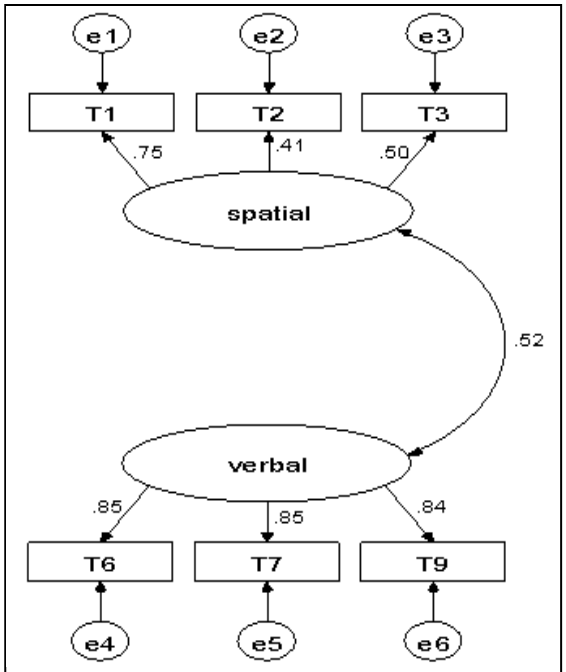

Figure 5. AMOS Output Model with Error Variances Estimated from the Data

Model #4 of 4: Using Previously Reported Reliabilities

To illustrate for these data the results when *score reliabilities from prior research reports are used in the analysis*, the six error variances (i.e., the measurement error variances iff the model is presumed to be correctly specified) in the Input Model were all constrained to reflect previously reported (Harman, 1976) Cronbach's alpha coefficients for the six variables: 0.756, 0.568, 0.544, 0.651, 0.754, and 0.870, respectively. Specifically, input error variances were constrained to equal the observed variance for a given variable times (1 - the previously reported Cronbach's alpha) for that variable. Thus, for example, the error variance for T1 in this analysis was set equal to (1 - 0.756) times 49.064, or 11.972. Our use of Cronbach's alpha in this illustration should not be taken to mean that the type of reliability estimated within SEM is necessarily alpha, or that SEM is limited to estimating only reliabilities that assume (a) measures are "tau-equivalent" and (b) errors are uncorrelated.

Figure 6 presents the Output Model parameter estimates for the standardized model. The interested reader can reproduce these results by applying the Appendix E AMOS syntax commands to the Appendix A data file. For this analysis, the Output Model estimated correlation between Spatial Ability and Verbal Ability was 0.46. The four unconstrained factor pattern coefficients were 0.59 and 0.62 for T2 and T3 on Spatial Ability, respectively, and 0.81 and 0.86 for T6 and T7 on Verbal Ability, respectively.

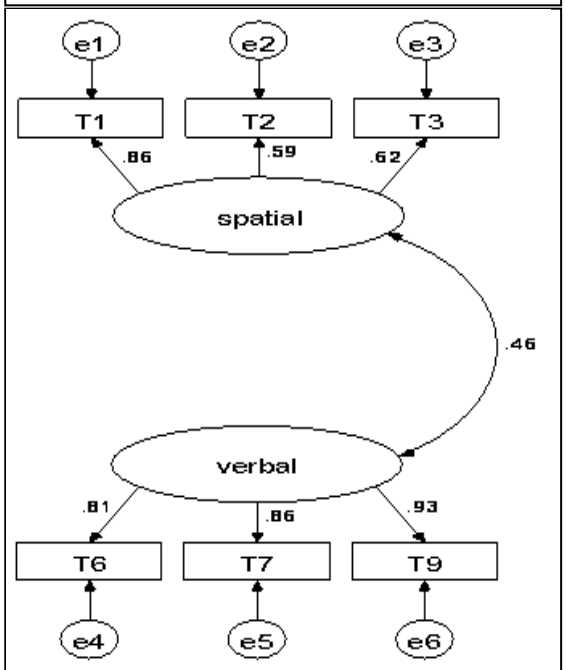

Figure 6. AMOS Output Model Using Previously Published Error Variances.

Discussion

The purposes of the present paper were to (a) explicate exactly where and how score reliability is estimated within structural equation modeling (SEM), and (b) illustrate exactly how changes in score reliability estimates can impact all the estimates in a given analysis. First, with respect to our first objective, we have noted that we are estimating the score reliabilities of our measured variables *when we are estimating the error variances of our measured variables*, even when the reliability calculations are not routinely provided in our outputs! Within the GLM, these estimates of error variances are *only* made within SEM.

Second, with respect to demonstrating the impacts of score reliabilities on all the parameters within our statistical analyses, whatever these analyses may be, clearly the estimated correlation coefficients

between Spatial Ability and Verbal Ability as well as the four estimated factor pattern coefficients (i.e., T2 and T3 on Spatial Ability, and T6 and T7 on Verbal Ability) all differed across the four sets of reliability estimates. For example, across the four analyses the interfactor correlation coefficients were estimated as 0.38, 0.80, 0.52, and 0.46, respectively.

The largest estimate of the correlation between the Spatial Ability and Verbal Ability factors (i.e., r = 0.80) occurred for the model in which the analysis (i.e., #2) assumed that all six score reliabilities were near zero. In effect, the analysis therefore inflated the estimated factor correlation from the estimate that would have occurred had better score reliability been presumed.

This correction is analogous to the classical correction of a correlation coefficient for imperfect score reliability (Spearman, 1910):

$$r_{X'Y'} = r_{XY} / [r_{XX} (r_{YY})]^{.5}$$

where r_{XY} is the uncorrected correlation coefficient between X and Y, r_{XX} is the reliability coefficient for the X scores, r_{YY} is the reliability coefficient for the Y scores, and $r_{X'Y'}$ is the correlation coefficient corrected for imperfect score reliability. For example, for r_{T1xT2} = 0.297, r_{T1xT1} = 0.756, and r_{T2xT2} = 0.568, we obtain $r_{X'Y'}$:

$$0.297 / [0.756 (0.568)]^{.5}$$
$$0.297 / [0.429]^{.5}$$
$$0.297 / 0.655 = 0.454.$$

In closing, we summarize our arguments by stating the following precepts:

1. Score reliability affects our estimates in all statistical analyses within the GLM (i.e., SEM and the analyses subsumed by SEM), whether these reliabilities are estimated within statistical analyses, which they are *not* in *t* tests, ANOVA, ANCOVA, Pearson *r*, Hotelling's T^2, MANOVA, MANCOVA, descriptive discriminant analysis, or CCA, or are estimated with SEM analyses. Even in SEM, score reliability estimates do affect our overall fit statistics, and so the quality of our measurement error estimates is important even in SEM.
2. Score reliabilities should be computed for the data in hand whenever GLM methods other than SEM are being used, so that measured variables with poor reliabilities can be excluded from analyses, even though roughly only one-third of published articles report reliabilities for the scores being analyzed (Vacha-Haase, Henson & Caruso, 2002; Willson, 1980), and more than half the published articles never even mention score reliability!

References

Arbuckle, J. L. (2007). *AMOS™ 16.0 user's guide*. Spring House, PA: Amos Development Corporation.

Bagozzi, R.P., Fornell, C., & Larcker, D.F. (1981). Canonical correlation analysis as a special case of a structural relations model. *Multivariate Behavioral Research, 16*, 437-454. doi:10.1207/s15327906mbr1604_2

Cohen, J. (1968). Multiple regression as a general data-analytic system. *Psychological Bulletin, 70*, 426-433. doi:10.1037/h0026714

Cronbach, L. J., Gleser, G. C., Nanda, H., & Rajaratnam, N. (1972). *The dependability of behavioral measurements: Theory of generalizability for scores and profiles*. New York, NY: Wiley.

DeShon, R. P. (1998). A cautionary note on measurement error corrections in structural equation models. *Psychological Methods, 3*, 412-423. doi:10.1037/1082-989X.3.4.412

Fan, X. (1997). Canonical correlation analysis and structural equation modeling: What do they have in common? *Structural Equation Modeling, 4*, 65-79. doi:10.1080/10705519709540060

Fisher, R. A. (1936). The use of multiple measurements in taxonometric problems. *Annals of Eugenics, 7,* 179-188.

Graham, J. M. (2008). The General Linear Model as structural equation modeling. *Journal of Educational and Behavioral Statistics, 33*, 485-506. doi:10.3102/1076998607306151

Harman, H. H. (1976). *Modern factor analysis* (3rd rev. ed.). Chicago: The University of Chicago Press.

Henson, R.K., & Thompson, B. (2002). Characterizing measurement error in scores across studies: Some recommendations for conducting "Reliability Generalization" (RG) studies. *Measurement and Evaluation in Counseling and Development, 35*, 113-127.

Holzinger, K. J., & Swineford, F. (1939). *A study in factor analysis: The stability of a bi-factor solution* (No. 48). Chicago, IL: University of Chicago.

Kieffer, K. M., Reese, R. J., & Thompson, B. (2001). Statistical techniques employed in AERJ and JCP articles from 1988 to 1997: A methodological review. *Journal of Experimental Education, 69*, 280-309. doi:10.1080/00220970109599489

Knapp, T. R. (1978). Canonical correlation analysis: A general parametric significance testing system. *Psychological Bulletin, 85*, 410-416. doi:10.1037/0033-2909.85.2.410

Shaw, D., Kopriva, R., & Tracz, S. (1993). Equations which include the reliability of the dependent variable for estimating the power of a two group ANOVA design. *Multiple Linear Regression Viewpoints, 20*(1), 1-5. http://mlrv.ua.edu/1993/MLRV_SPRING1993_VOLUME20__1.PDF

Spearman, C. E. (1904). The proof and measurement of association between two things. *American Journal of Psychology, 15*, 72-101. doi:10.2307/1412159

Spearman, C. E. (1910). Correlation calculated from faulty data. *British Journal of Psychology, 3*, 271-295.

Thompson, B. (1984). *Canonical correlation analysis: Uses and interpretations*. Thousand Oaks, CA: Sage.

Thompson, B. (1991). A primer on the logic and use of canonical correlation analysis. *Measurement and Evaluation in Counseling and Development, 24*, 80-95.

Thompson, B. (2000). Ten commandments of structural equation modeling. In L. Grimm & P. Yarnold (Eds.), *Reading and understanding more multivariate statistics* (pp. 261-284). Washington, DC: American Psychological Association.

Thompson, B. (2003a). A brief introduction to Generalizability Theory. In B. Thompson (Ed.), *Score reliability: Contemporary thinking on reliability issues* (pp. 43-58). Newbury Park, CA: Sage.

Thompson, B. (2003b). Understanding reliability and coefficient alpha, really. In B. Thompson (Ed.), *Score reliability: Contemporary thinking on reliability issues* (pp. 3-23). Newbury Park, CA: Sage.

Thompson, B. (2004). *Exploratory and confirmatory factor analysis: Understanding concepts and applications*. Washington, DC: American Psychological Association.

Thompson, B. (2006a). *Foundations of behavioral statistics: An insight-based approach*. New York: Guilford.

Thompson, B. (2006b). Research synthesis: Effect sizes. In J. Green, G. Camilli, & P. B. Elmore (Eds.), *Handbook of complementary methods in education research* (pp. 583-603). Washington, DC: American Educational Research Association.

Thompson, B. (2007). Effect sizes, confidence intervals, and confidence intervals for effect sizes. *Psychology in the Schools, 44*, 423-432.

Thompson, B., & Vacha-Haase, T. (2000). Psychometrics is datametrics: The test is not reliable. *Educational and Psychological Measurement, 60*, 174-195. doi:10.1177/0013164400602002

Vacha-Haase, T. (1998). Reliability generalization: Exploring variance in measurement error affecting score reliability across studies. *Educational and Psychological Measurement, 58*, 6-20.

Vacha-Haase, T., Henson, R. K., & Caruso, J. (2002). Reliability Generalization: Moving toward improved understanding and use of score reliability. *Educational and Psychological Measurement, 62*, 562-569. doi:10.1177/0013164402062004002

Wang, Z., & Thompson, B. (2007). Is the Pearson r^2 biased, and if so, what is the best correction formula? *Journal of Experimental Education, 75*, 109-125.

Wilkinson, L., & APA Task Force on Statistical Inference. (1999). Statistical methods in psychology journals: Guidelines and explanations. *American Psychologist, 54*, 594-604. doi:10.1037/0003-066X.54.8.594

Willson, V. L. (1980). Research techniques in *AERJ* articles: 1969 to 1978. *Educational Researcher, 9*(6), 5-10. doi:10.3102/0013189X009006005

Wright, S. (1921). Correlation and causality. *Journal of Agricultural Research, 20*, 557-585.

Wright, S. (1934). The method of path coefficients. *Annals of Mathematical Statistics, 5*, 161-215.

Zientek, L. R., & Thompson, B. (2009). Matrix summaries improve research reports: Secondary analyses using published literature. *Educational Researcher, 38*, 343-352.

Send correspondence to: Z. Ebrar Yetkiner or Bruce Thompson
Texas A & M University
Email: ebraryetkiner@yahoo.com or bruce-thompson@tamu.edu

APPENDIX A

ASCII Text Data File ("H_S.TXT") for the Models Tested

```
rowtype_,varname_,T1,T2,T3,T6,T7,T9
n, 301, 301, 301, 301, 301, 301
corr,T1,1.000
corr,T2, 0.297346,  1.000
corr,T3, 0.365293,  0.237982,  1.000
corr,T6, 0.372706,  0.152930,  0.211914,  1.000
corr,T7, 0.293444, 0.139387, 0.173404, 0.733170,  1.000
corr,T9, 0.356770, 0.192532, 0.238544, 0.704480,  0.719956,  1.000
stddev,, 7.004593, 4.709802, 2.830302, 3.492349, 5.161889, 7.669222
mean,, 29.614618, 24.352159, 14.229236, 9.182724, 17.362126, 15.299003
```

APPENDIX B

Syntax for the Model Where Error Variances = 0.01

```
Sub Main
Dim sem As New AmosEngine

sem.TextOutput
sem.Standardized
sem.Smc
sem.Mods

sem.BeginGroup "H_S.TXT"

sem.Structure "T1 =   (1) spatial + (1) e1"
sem.Structure "T2 =       spatial + (1) e2"
sem.Structure "T3 =       spatial + (1) e3"
sem.Structure "T6 =       verbal  + (1) e4"
sem.Structure "T7 =       verbal  + (1) e5"
sem.Structure "T9 =   (1) verbal  + (1) e6"
sem.Structure "spatial <--> verbal"
sem.Structure "e1 (0.01)"
sem.Structure "e2 (0.01)"
sem.Structure "e3 (0.01)"
sem.Structure "e4 (0.01)"
sem.Structure "e5 (0.01)"
sem.Structure "e6 (0.01)"

End Sub
```

APPENDIX C

Syntax for the Model Where Error Variances are Set Equal
to the Total Variances of Observed Variables

```
Sub Main
Dim sem As New AmosEngine

sem.TextOutput
sem.Standardized
sem.Smc
sem.Mods

sem.BeginGroup "H_S.TXT"

sem.Structure "T1 =   (1) spatial + (1) e1"
sem.Structure "T2 =       spatial + (1) e2"
sem.Structure "T3 =       spatial + (1) e3"
sem.Structure "T6 =       verbal  + (1) e4"
sem.Structure "T7 =       verbal  + (1) e5"
sem.Structure "T9 =   (1) verbal  + (1) e6"
sem.Structure "spatial <--> verbal"
sem.Structure "e1 (49.06431894)"
sem.Structure "e2 (22.18223699)"
sem.Structure "e3 (8.010609081)"
sem.Structure "e4 (12.19650055)"
sem.Structure "e5 (26.64509413)"
sem.Structure "e6 (58.81696567)"

End Sub
```

APPENDIX D

Syntax for the Model Where Error Variances are Set Free to be Estimated in the Analysis

```
Sub Main
Dim sem As New AmosEngine

sem.TextOutput
sem.Standardized
sem.Smc
sem.Mods

sem.BeginGroup "H_S.TXT"

sem.Structure "T1 =   (1) spatial + (1) e1"
sem.Structure "T2 =       spatial + (1) e2"
sem.Structure "T3 =       spatial + (1) e3"
sem.Structure "T6 =       verbal  + (1) e4"
sem.Structure "T7 =       verbal  + (1) e5"
sem.Structure "T9 =   (1) verbal  + (1) e6"
sem.Structure "spatial <--> verbal"

End Sub

sem.Structure "e6 (7.646205537)"

End Sub
```

APPENDIX E

Syntax for the Model Where Error Variances are Set Equal to (1 - Cronbach's Alpha) * Total Variances of Observed Variables

```
Sub Main
Dim sem As New AmosEngine

sem.TextOutput
sem.Standardized
sem.Smc
sem.Mods

sem.BeginGroup "H_S.TXT"

sem.Structure "T1 =   (1) spatial + (1) e1"
sem.Structure "T2 =       spatial + (1) e2"
sem.Structure "T3 =       spatial + (1) e3"
sem.Structure "T6 =       verbal  + (1) e4"
sem.Structure "T7 =       verbal  + (1) e5"
sem.Structure "T9 =   (1) verbal  + (1) e6"
sem.Structure "spatial <--> verbal"
sem.Structure "e1 (11.97169382)"
sem.Structure "e2 (9.582726379)"
sem.Structure "e3 (3.652837741)"
sem.Structure "e4 (4.256578693)"
sem.Structure "e5 (6.554693156)"
```

Prediction Accuracy: A Monte Carlo Comparison of Several Methods in the Continuous Variable Case

Holmes Finch **Jocelyn Holden**
Ball State University

Ordinary least squares (OLS) regression is by far the most popular method for prediction that is used in the social sciences. While this approach can provide accurate prediction in many contexts, it may not be optimal in every case, particularly for modeling non-linear relationships, unless the user already has a sense for what these are likely to be and includes them in the analysis. On the other hand, a number of alternative methods have emerged in the last 20 years that may prove to be more useful in certain circumstances. The current simulation study sought to investigate the relative performance of OLS with several of these alternatives for predicting a continuous outcome variable for linear and non-linear relationships. The results demonstrated that for standard linear relationships, and those with a single interaction, OLS provides relatively accurate predictions, although it is rarely the optimal method. On the other hand, for more complex relationships, OLS does not perform as well as several modeling alternatives. The discussion provides researchers with some guidance regarding which methods might be optimal to use under varying conditions.

Ordinary least squares regression (OLS) has historically been the method employed when prediction is the goal of a study. In 1977, Hill and Holland stated "With the advent of modern computers, not only are complex regression analyses common, they are routinely done in many fields, and this increase in the use of regression has produced a need for refinements in this technology which will alert users to possible problems," (Hill & Holland, 1977, p. 828).

This statement remains accurate some 30 years later. A review of the PsycInfo database for articles published in the last 10 years shows that when prediction is the goal of a social science study, OLS is by far the most popular method of choice. The OLS model takes the form:

$$Y_i = \beta_0 + \sum_{j=1}^{J} \beta_j x_j \quad (1)$$

where β_0 = the model intercept and β_j = slope relating variable x_j to Y_i. The model coefficients are determined using the least squares criteria, which seeks to minimize the sum of the squared residuals. While OLS has a linear form by default, if the researcher believes that there exist interactions among the independent variables, and/or non-linear relationships between individual *x*'s and the response, these can be built into the model explicitly. This does require, however, that the user have some predetermined notion as to the nature of these non-linear terms.

Although OLS continues to be the most common method of prediction, investigation of alternative prediction tools is not new. Over the last 30 years, many different techniques have been proposed and compared with OLS in search of a more powerful, more accurate (Dempster, Schatzoff, & Wermuth, 1977; Park & Kim, 1997) or more robust (Hill & Holland, 1977; Hussain & Sprent, 1983) model. Early conceptualizations of the problem lead to developments in estimation and modeling of nonparametric regression techniques (Hussain & Sprent, 1977; Park & Kim, 1997). Others have argued that use of simpler models incorporating clinical judgment, correlation coefficients, or unit weights rather than OLS regression weights can improve the accuracy and generalizability of prediction models (Dana & Dawes, 2004; Dawes, 1979; Dawes & Corrigan, 1977; Waller & Jones, 2009).

More recent advances in prediction modeling have introduced new methods not based on linear models at all, but rather incorporating (automatically in some cases) more complex non-linear relationships between dependent and independent variables. In particular, models such as neural networks (NNET), Classification and Regression Trees (CART), Generalized Additive Models (GAMs), Multivariate Adaptive Regression Splines (MARS), and boosting (Kuhnert, Mengersen, & Tesar, 2003) are increasing in popularity, especially in the fields of Ecology, Business, and Medicine. These newer methods provide more flexible alternatives to OLS by automatically identifying more complex model forms to fit the response variable, which are not easily incorporated into standard linear regression methodologies. Despite their use in a variety of disciplines, however, they are seldom seen in social science research. The goal of the current Monte Carlo study was to compare the relative ability of several of these alternatives for model prediction with the standard OLS methodology under a variety of

data conditions. The manuscript begins with a description of the prediction methods to be studied here, which were selected because of their successful use in other areas. Next, the methodology of this study is described, followed by the Monte Carlo results. Finally, we discuss the results of this study and provide recommendations for practice.

Model Descriptions

Classification and Regression Trees (CART)

CART (Breiman, Friedman, Olshen, & Stone, 1984; Williams, Lee, Fisher, & Dickerman, 1999) arrives at predicted values for *Y* given a set of predictors by iteratively dividing individual members of the sample into ever more homogeneous groups, or nodes, based on values of the predictor variables. It can be thought of as a nonparametric approach in that there are no assumptions regarding the underlying population from which the sample is drawn nor the form of the model linking the dependent and independent variables. CART begins by placing all subjects into one node, and then searches the set of predictors to find the value of one of those by which it can divide the observations into two new nodes, whose values on *Y* are as homogeneous as possible. For each of these new nodes, the predictors are once again searched for the optimal split by which the subjects can be further divided into ever more homogeneous nodes, where homogeneity is always based on similarity of values on *Y*. This division of the data continues until a predetermined stopping point is reached, when further splits do not appreciably reduce the heterogeneity of the resulting nodes. At this point, the tree is complete and values of *Y* for new observations can be obtained using the splits developed with this training sample.

Neural Networks (NNET)

Another prediction method included in this study is Neural Networks (NNET) (e.g., Marshall & English, 2000). See Garson (1998) for a more thorough description of the method designed especially for social scientists. As with CART, NNETs have not been used extensively in the social sciences, and yet would seem to offer some potential advantages over OLS in certain circumstances. NNETs find relationships between *Y* and a set of predictor variables by using a search algorithm that examines a large number of subsets of the predictors, and interactions among them, in conjunction with various weights. These interactions and higher order versions of the predictor variables are selected so as to minimize the common least squares criterion briefly described above in the context of OLS, and may involve multiple weights applied to virtually any combination of the predictors. In order to reduce the likelihood of finding locally optimal results that will not generalize beyond the original (training) sample, random changes to the subsets not based on model fit are also made. This method of ascertaining fit is known as back-propagation, where the difference between actual and predicted outputs is used to find optimal weight values. This is one of the most commonly used approaches in NNET applications (Garson, 1998). There are a number of NNET models available for use. Perhaps the most common of these (and the one utilized in this study) is the feed-forward back propagation network with one hidden layer. This particular NNET architecture uses the least squares minimization method e in order to obtain weights for the inputs, or predictor variables, and includes what is known as a hidden layer, which is analogous to one or more interactions in the more familiar regression context (Garson, 1998). It should be noted, however, that nodes in this hidden layer can be much more complicated than the interactions one might see in a standard linear model, involving complex combinations of the weighted predictor variables (Schumacher, Robner & Vach, 1996). Finally, the inputs and hidden nodes are used in conjunction with the weights in order to obtain the predicted value of *Y*.

One of the purported strengths of the NNET approach is that it can identify complex interactions among the predictor variables in the hidden layer that other approaches may ignore (Marshall & English, 2000). For example, whereas in regression it is common to express the interaction of two predictors as their product, or to square or cube a single variable if the relationship with the response is believed not to be linear, a NNET will create hidden nodes as weighted products of perhaps several variables, thus allowing the hidden nodes to be influenced by the predictors in varying degrees.

A potential problem with using NNET's is the possibility of substantial overfitting of the data (Schumacher, Robner, & Vach, 1996). In other words, the weights selected for each variable and each hidden layer may be so closely linked to the original, or training sample that the results are non-generalizable to a wider population. Overfitting is typically identified through the use of cross-validation of results obtained with the training sample (i.e., applying the weights as identified by the training sample

to a second sample ostensibly drawn from the same population). In order to combat this problem, most NNET models apply what is called decay, or weight decay, which penalizes (i.e., reduces) the largest weights found in the original NNET analysis, in effect assuming that very large weights are at least partially driven by random variation unique to the observed data.

Multivariate Adaptive Regression Splines (MARS)

The MARS model is an extension of the linear regression model in which non-linear relationships and interactions are modeled automatically. It takes the form:

$$f(x) = \beta_0 + \beta_m h_m(x) \quad (2)$$

where β_0 = the model intercept and β_m = coefficient and $h_m(x)$ = function of the independent variable x from a set of basis functions. These basis functions (sometimes known as hinge functions) are piecewise linear and expressed as max(0, x-t) and max (t-x,0), where t is the knot, or the location where the relationship between a predictor and Y changes direction (Hastie, Tibshirani, & Friedman, 2001). For a given $h_m(x)$, the model coefficients are estimated through ordinary least squares, minimizing the residual sum of squares.

MARS begins building a predictive model using a forward stepwise methodology, in which the first step involves the inclusion only of β_0. The algorithm then proceeds to add a basis function to the model in each step, selecting the $h_m(x)$ that provides the greatest reduction in the sum of squared residuals. At each step, the newly added basis function will include a term already in the model multiplied by a new hinge function. When deciding which new basis function to add to the existing model, MARS searches across all terms currently in the model, all of the independent variables included in the analysis and each possible value for the independent variables, in order to select the knot for the new basis function. The model building continues until the change in sum of squared residuals becomes very small when a new term is entered, or until a maximum model size (set by the user) is reached.

As with NNET, the MARS model will likely be overfit to the data. Therefore, a stepwise backward deletion procedure is included, in which the least important (from a statistical sense) term is removed at each step. In order to identify the optimal subset of model terms, the Generalized Cross-Validation (GCV) criterion is minimized. The GCV is a function of the sum of squared residuals, adjusted by the number of parameters in the model and the sample size.

The primary advantage of the MARS model building strategy is its ability to work well locally in the function space (Hastie, Tibshirani, & Friedman, 2001). Specifically, the use of the basis functions described above allows for the modeling of interactions only in the range of data for which two such functions have a non-zero value. Thus, unlike the more general polynomial terms in regression, the entire data space is not required to take a common non-linear functional form.

Generalized Additive Models (GAM)

GAM's are a class of very flexible models that allow for the linking of Y with one or more independent variables, using a wide variety of smoothing functions. The GAM itself takes the form:

$$Y = \beta_0 + \sum_{j=1}^{p} f_j(x_j) \quad (3)$$

where β_0 = the model intercept and $f_j(x_j)$ = a smoothing function for independent variable x_j. Each of the functions is fit using a scatterplot smoothing technique such as cubic splines or a Kernel smoother, with the goal of minimizing the penalized sum of squares criterion to identify the optimal GAM for a given problem. The penalized sum of squares (PSS) is based on the standard sum of squares (i.e. the difference between the actual and predicted values of the response variable), with a penalty for model complexity applied.

The GAM algorithm works in an iterative fashion, beginning with the setting of β_0 to the mean of Y. Subsequently, a smoothing function is applied to each of the independent variables in turn, minimizing the PSS. The iterative process continues until the smoothing functions stabilize, at which point final model parameter estimates are obtained. The most common smoothing function used with GAM's (and the one used in this study) is the cubic spline (Simonoff, 1996). As is the case for several of the methods described in this paper, overfitting of the data can be a problem with GAMs. Therefore, it is recommended that the number of smoothing parameters be kept relatively small, and that cross-validation

be used to ensure that the resulting model is generalizable to other samples from the target population (Wood, 2006).

Boosting (BOOST)

Boosting refers to a collection of machine learning algorithms that attempt to improve prediction of *Y* by combining the predictions from a set of weak predictor variables in order to obtain a single strong prediction (Freund & Schapire, 1997). These ideas were first introduced by Tukey (1977) in a method he referred to as "twicing," and subsequently other researchers have expanded on this idea. Boosting was originally developed for use in predicting a dichotomous outcome variable, but has since been generalized to the regression context with continuous responses (see Buhlmann & Hothorn, 2007 for a discussion of this history). Of these extensions, one of the more popular for use in the regression context with a continuous outcome is L_2 boosting, which consists of 5 steps:

1) Initialize the function linking the response with the predictors using (typically) OLS.
2) Compute residuals from step 1, $U_i = Y_i - \hat{F}_m(x_i)$ where $\hat{F}_m(x_i)$ is the regression function obtained in 1) applied to the predictor variables x_i at iteration *m*.
3) Fit the residuals U_i using the independent variables x_i with OLS.
4) Use the fitted values obtained in 3), denoted $\hat{f}_{m+1}$, to update the original fit $\hat{F}_m(x_i)$ to obtain $\hat{F}_{m+1} = \hat{F}_m + \hat{f}_{m+1}$.
5) Repeat steps 1 through 4 until the final iteration is reached.

The number of iterations to be used with the boosting algorithm is essentially a tuning parameter. A variety of approaches for determining when to stop the boosting algorithm have been recommended, with perhaps the most recent (Buhlmann & Hothorn, 2007) being the use of the minimum value for Akaike's Information Criterion (AIC). Therefore, a researcher may elect to use a large number of *m* iterations, and then review the resultant AIC values, selecting the model that corresponds with the smallest of these, which was the approach used in the current study.

L_2 boosting will typically lead to complex model forms as the number of iterations (and thus residual functions) increases. For this reason, it is generally recommended that the original regression model used to predict *Y* be fairly simple, consisting of a relatively small number of predictor terms (Buhlmann & Yu, 2003). Finally, it should be noted that while linear regression is quite often the model to which the boosting algorithm is applied, it is entirely possible to use smoothing splines or other functions to relate the response variable to the predictors and then apply the boosting algorithm (Buhlmann & Hothorn, 2007).

Correlation weights (CORR)

An alternative approach to deriving prediction equations that has been suggested is the use of weights based on the zero order correlation coefficients between the individual standardized predictors and the standardized *Y* (Dana & Dawes, 2004). Using this approach, the model for predicting *Y* with three predictors, x_1, x_2, x_3 would be expressed as:

$$Y_i = r_{x1Y}(x_1) + r_{x2Y}(x_2) + r_{x3Y}(x_3) \quad (4)$$

where r_{xjY} = Pearson correlation relating x_j to *Y*. Dana and Dawes (2004) note that both Mark (1966) and Goldberg (1972) reported positive results when using these correlations as opposed to the standard OLS regression weights.

Previous comparisons of prediction methods

There has been some work examining the ability of these various methods to accurately predict outcome variables, both categorical and continuous, although none of these prior studies have simultaneously compared all of these methods with one another. Much of this work has involved comparing two or three of these methods with one another using a single set of real data for which the actual population parameter values are not known. In terms of the prediction of dichotomous outcomes, MARS (Munoz & Felicisimo 2004) and neural networks (Finch & Schneider, 2007; Reibnegger, Weiss, Werner-Felmayer, Judmaier, & Wachter, 1991) have been found to achieve higher prediction accuracy than other more common techniques (logistic regression, discriminant function analysis, principle

components regression). In terms of prediction of continuous outcomes, again, MARS (Moisen & Frescino, 2002), and neural networks (Nguyen & Cripps, 2001; Razi & Athappilly, 2005; Smith & Mason; 1997) have been found to demonstrate better sensitivity and accuracy than OLS when the underlying model was non-linear. In addition GAM (Moisen & Frescino, 2002) and CART models (Lee & Jin, 2006; Razi & Athappily, 2005) have also been shown to provide better estimates than traditional regression methods for continuous outcomes.

Often the choice of prediction rests on characteristics of the study and data used. Lee & Jin (2006) demonstrate that CART models can provide better prediction than Poisson and zip regression in circumstances of zero-inflated count data. GAMs and MARS demonstrate large gains in accuracy and sensitivity over OLS regression when data is simulated, however when real data is used the gains are much less prominent (Moisen & Frescino, 2002). MARS is also especially powerful when detecting higher order interactions (Munoz & Felicisimo, 2004) and generally produces the fastest solutions of all the newer models (Moisen & Frescino, 2002). When complicated data structures (Reibnegger et al., 1991; Smith & Mason, 1997) are encountered, especially with large sample sizes, (Nguyen & Cripps, 2006) neural networks can be particularly powerful. In addition, neural networks and CART models demonstrate better accuracy than traditional regression models when categorical predictor variables are incorporated (Razi & Athappily, 2005). Again, it is important to note that many of these prior studies relied only on single real datasets for which the underlying model was not known, and none of these simultaneously compared all of the methods examined in the current work. Thus, the goal of this study is to compare the prediction accuracy of OLS with CART, NNET, MARS, GAM, CORR and BOOST for a continuous outcome under a variety of simulated model and data conditions.

Methods

Data Generation

Generation of simulated data (1000 replications for each condition) based on conditions contained in Table 1, and the subsequent prediction analyses were performed using the R statistical software program (R Development Core Team, 2007). An example of the R code used to generate the simulated data and run the analyses appears in the appendix. Correlations among the predictor variables were drawn from those found among WAIS-III subscales, which appeared in Waller and Jones (2009). These correlations, which can be seen in the R code appearing in the appendix of this manuscript, range from 0.36 to 0.76 and might be thought of as falling in the moderate to large range. Controlling of inter-predictor interactions was accomplished by defining simple-complex functions for the outcome variable (equations appear below). Model complexity took three forms, which appear below: (1) Simple additive linear model with no interactions, (2) Single interaction condition, where the outcome variable was defined as a linear combination of the predictors plus a two-way interaction between x_1 and x_2, and (3) Complex interaction condition for which the outcome variable was defined as a linear combination of the predictors plus a two-way interaction, a squared term and a cubed term. Once the base function was defined, in order to control for predictor strength, slopes for each variable were defined at one of three conditions 0.2, 0.5, or 0.8. For the sake of simplicity all predictors had the same strength for each replication.

Table 1. Simulation Conditions

Variable	Levels
Sample Size	50, 100, 500, 1000
Predictor Strength	0.2, 0.5, 0.8
Predictor Interactions	none, simple, complex
Noise	yes, no
Prediction Method	OLS, CART, GAM, MARS, COR, BOOST, NNET

$$
\begin{aligned}
&1.\quad y = x_1 + x_2 + x_3 + x_4 \\
&2.\quad y = x_1 + x_2 + x_3 + x_4 + x_1 * x_2 \\
&3.\quad y = x_1 + x_2 + x_3 + x_4 + x_1 * x_2 + x_3^2 + x_4^3
\end{aligned}
\tag{5}
$$

The noise condition refers to whether or not there were extraneous variables included in the predictive model for the sample that were not included in the simulated population model. This would be akin to a researcher including variables in their predictive equation that they believed were related to the outcome, but which in actuality were not related in the population. The no noise condition used only the four variables associated with the outcome in the population to predict the outcome variable, while the noise condition included an additional four extraneous variables to predict the outcome. More specifically, the

noise condition was simulated by including independent variables in the prediction model for each replication whose coefficient with the dependent variable was simulated from a population value of 0. As an example, the population equation from which the data were generated would take the form: $y=x_1+x_2+x_3+x_4$, while the predictive models would take the form $y=x_1+x_2+x_3+x_4+x_5+x_6+x_7+x_8$. The noise variables in this case are x_5-x_8.

It is well known (Berk, 2008) that predictions for the training data will be unrealistically accurate, or overfit. Therefore, in order to appropriately test the accuracy of each method, it is necessary to first estimate a prediction model with the training data, and then test the prediction accuracy with a separate cross-validation sample. Thus, for each simulation replication two datasets were generated: 1) training data with which the sample model was generated, and 2) cross-validation data for which predictions using the model derived from 1) were made. Both samples were simulated from the same population, for each replication in this study.

Analyses. Once data were generated, predictions of the outcome were made using the seven different prediction methods previously discussed. For each replication and prediction method, the root mean square error (RMSE) for the cross-validation sample and the correlation between the actual values and the predicted values from the cross-validation sample were saved. These, or similar outcomes have been used in prior work (Moisen & Frescino, 2002; Razi & Athappily, 2005) of which this study is meant to be an extension. Specifically, the RMSE reflects overall prediction accuracy, accounting for both estimation bias and variation. In addition, the correlation between predicted and actual values provides information regarding the pattern of prediction accuracy, above and beyond that given by RMSE. The average RMSE and correlations across the thousand iterations are reported in the results. For each of these methods the default settings in the R software package were used, values of which can be found in the software documentation. For CART there was no pruning of the trees conducted, an issue that is discussed below. With respect to NNET, several different models were run with respect to the number of hidden nodes including 1, 2, 3, 4, 5, and 15 such nodes. For each combination of conditions the model with the lowest RMSE value was selected in order to simplify reporting.

Results

RMSE

In order to determine which of the manipulated effects were significantly related to the RMSE for the predictions, a full factorial ANOVA model was used, in which the independent variables were method of prediction, model complexity, sample size, strength of the relationship between the independent variables and the response, and whether noise was included in the model. Two 4-way interactions among these manipulated factors were found to be significant: method by sample size by model complexity by strength (p=0.018, η^2=0.738) and method by noise by strength by model complexity (p=0.002, η^2=0.819). All other significant interactions and main effects were subsumed under these two interactions and thus will not be discussed further. Table 2 includes the RMSE values for each method by model complexity, sample size and strength of the predictors. It is important to note that for each replication, the predictions discussed here were made on the cross-validation sample, while the models used to make the predictions were based on the training sample.

No-Interaction Model

An examination of these results reveals that for the simplest model (no interactions among predictors), CORR and BOOST methods consistently produced the most accurate estimates, reflected in the lowest RMSE values, regardless of sample size or strength of relationship. For this condition, OLS, GAM and MARS performed very similarly, and always somewhat less well than CORR and BOOST methods. As with these two top performers, the sample size and strength of relationship appears to have had little impact on RMSE for OLS, GAM, and MARS. The NNET method was somewhat more affected by sample size for the least complex model condition, so that for smaller samples it had a higher RMSE than OLS, GAM, or MARS, but for a sample of 1000 it produced very comparable results to these other approaches. Finally, in this simplest model case, CART uniformly had the highest RMSE values, indicating that it produced the least accurate predictions for the cross-validation sample regardless of sample size or strength of the predictors. Unlike the other methods studied here, the accuracy of CART was impacted by the strength of the predictors, such that the RMSE increased concomitantly with the

strength of the relationships. In addition, as the sample size increased, the RMSE for CART decreased, indicating that with larger sample sizes CART predictions were somewhat more accurate, although they were never as accurate as those of the other methods.

Table 2. RMSE by Prediction Method, Level of Interaction, Sample Size and Predictor Strength

Interaction	N	Strength	OLS	CART	GAM	MARS	CORR	BOOST	NNET
Complex	50	.2	0.502	0.772	0.196	0.384	2.410	2.410	0.512
		.5	1.248	1.938	0.367	0.900	6.151	6.151	1.212
		.8	2.002	3.022	0.556	1.432	9.751	9.751	2.075
	100	.2	0.448	0.623	0.152	0.249	2.434	2.434	0.331
		.5	1.116	1.509	0.242	0.555	6.093	6.093	.897
		.8	1.750	2.397	0.348	0.877	9.753	9.753	1.528
	500	.2	0.382	0.434	0.120	0.147	2.442	2.442	0.219
		.5	0.987	1.002	0.137	0.238	6.173	6.173	0.550
		.8	1.510	1.645	0.177	0.382	9.720	9.720	1.001
	1000	.2	0.385	0.398	0.117	0.138	2.435	2.435	0.206
		.5	0.926	0.983	0.134	0.224	6.089	6.089	0.451
		.8	1.452	1.544	0.156	0.305	9.767	9.767	0.673
Simple	50	.2	0.111	0.321	0.122	0.143	1.054	1.054	0.178
		.5	0.111	0.819	0.174	0.231	2.632	2.632	0.257
		.8	0.111	1.278	0.232	0.322	4.213	4.213	0.381
	100	.2	0.115	0.290	0.124	0.133	1.058	1.058	0.137
		.5	0.116	.654	0.142	0.163	2.638	2.638	0.161
		.8	0.117	1.078	0.172	0.216	4.219	4.219	0.195
	500	.2	0.112	0.249	0.111	0.113	1.052	1.052	0.112
		.5	0.117	0.530	0.116	0.123	2.636	2.636	0.123
		.8	0.128	0.867	0.124	0.139	4.217	4.217	0.146
	1000	.2	0.115	0.241	0.116	0.117	1.054	1.054	0.117
		.5	0.119	0.523	0.116	0.121	2.638	2.638	0.128
		.8	0.132	0.802	0.120	0.131	4.215	4.215	0.148
None	50	.2	0.112	0.313	0.121	0.126	0.079	0.079	0.151
		.5	0.110	0.734	0.119	0.118	0.078	0.078	0.152
	100	.8	0.112	1.145	0.119	0.116	0.080	0.080	0.156
		.2	0.114	0.269	0.116	0.118	0.080	0.080	0.127
		.5	0.114	0.619	0.116	0.117	0.080	0.080	0.129
	500	.8	0.114	0.942	0.117	0.117	0.080	0.078	0.129
		.2	0.116	0.229	0.116	0.117	0.081	0.081	0.117
		.5	0.110	0.503	0.110	0.111	0.079	0.079	0.112
	1000	.8	0.116	0.770	0.116	0.116	0.079	0.079	0.117
		.2	0.112	0.217	0.113	0.112	0.080	0.080	0.113
		.5	0.112	0.473	0.112	0.112	0.079	0.079	0.113
		.8	0.113	0.738	0.113	0.113	0.080	0.080	0.114

Simple Interaction Model

For the simple interaction model (one interaction, between variables x_1 and x_2), the CORR and BOOST methods produced the highest RMSE values, indicating that they had the least accurate predicted values for the cross-validation samples. This result held true regardless of the sample size or strength of the relationships between the outcome and the predictors. In addition, while the sample size did not seem to have a demonstrable impact on the accuracy of these methods, their RMSE values did increase markedly as the strength of the predictors increased in value. In contrast, OLS, GAM, and MARS produced the lowest RMSE values in the middle complexity condition. Indeed, when the sample size was 50 or 100, OLS had the most accurate predictions, and was only slightly less accurate than GAM or MARS for *N* of 500 or 1000. Furthermore, the RMSE values for OLS increased concomitantly with the strength of the predictors for samples of 500 or 1000, while for GAM this impact of relationship strength was most noticeable for sample sizes of 50 or 100. The accuracy of MARS predictions was greater with larger samples, in which case the impact of relationship strength was mitigated somewhat. As was the case for the no interaction model, NNET had somewhat higher RMSE values than OLS, GAM or MARS, though this difference declined with increasing sample sizes. The pattern in terms of accuracy for CART was similar in the simple interaction condition as it was for the no interaction model, namely it had relatively high RMSE values across conditions (though not as high as those for the CORR prediction or BOOST), and it was somewhat more accurate for larger samples and for weaker correlations.

Table 3. RMSE by Prediction Method, Level of Interaction, Level of Noise and Strength of Relationship

Interaction	Noise	strength	OLS	CART	GAM	MARS	COR	BOOST	NNET
Complex	no	.2	0.429	0.548	0.146	0.231	2.429	2.429	0.251
		.5	1.060	1.322	0.216	0.479	6.109	6.109	0.615
		.8	1.664	2.151	0.305	0.756	9.730	9.730	1.041
	yes	.2	0.429	0.565	0.129	0.229	2.431	2.431	0.383
		.5	1.078	1.393	0.176	0.479	6.143	6.143	0.940
		.8	1.693	2.152	0.233	0.742	9.770	9.770	1.597
Simple	no	.2	0.113	0.278	0.118	0.127	1.055	1.055	0.126
		.5	0.116	0.627	0.135	0.157	2.635	2.635	0.138
		.8	0.123	0.999	0.161	0.202	4.215	4.215	0.166
	yes	.2	0.114	0.272	0.117	0.126	1.053	1.053	0.146
		.5	0.116	0.636	0.127	0.161	2.636	2.636	0.197
		.8	0.121	1.013	0.140	0.201	4.217	4.217	0.269
None	no	.2	0.114	0.254	0.117	0.119	0.080	0.080	0.123
		.5	0.111	0.591	0.113	0.114	0.078	0.078	0.121
		.8	0.113	0.915	0.115	0.115	0.079	0.079	0.123
	yes	.2	0.113	0.260	0.114	0.117	0.080	0.080	0.130
		.5	0.112	0.573	0.115	0.115	0.079	0.079	0.132
		.8	0.114	0.882	0.117	0.116	0.079	0.079	0.135

Complex Interaction Model

For the most complex underlying model, GAM produced the lowest RMSE values across methods examined in this study, followed by MARS. As was true for all methods in the complex interaction model case, except for CORR and BOOST, the RMSE values for GAM increased concomitantly with the strength of the predictors, and declined as the sample size increased. OLS and NNET performed similarly when the sample size was 50, but NNET had lower RMSE values than OLS for larger sample sizes, and the difference in performance between the two was magnified as *N* increased in size. In addition, both methods were less accurate when the strength of the predictors was greater. The worst performers by far in the complex interaction case were CORR and BOOST. As was true with some of

the other prediction methods, their RMSE values increased with the strength of the relationships, though their accuracy was not influenced by sample size. For this most complex model, CART had relatively high RMSE values, which increased with a greater strength of relationship between predictors and outcome, and which was lower for larger samples.

The second statistically significant interaction of the manipulated factors that was identified by the ANOVA involved the method of prediction by noise, by the strength of the predictors by model complexity. Table 3 contains the RMSE values for this interaction. Given the prior discussion of results relative to the model complexity and strength of relationships, the focus in this section is on the interaction of noise with these factors. A perusal of Table 3 shows that for the OLS, CART, MARS, CORR and BOOST approaches, the presence of noise variables in the model had little to no impact on the accuracy of predictions. On the other hand, for GAM the inclusion of noise variables resulted in slightly lower RMSE values for the most complex model condition, and for the middle model complexity case when the strength of relationship was 0.8. For the simplest model, the inclusion of noise variables had no impact on the RMSE for GAM, while it had the greatest effect on the performance of NNET for the most complex model, in which case the RMSE values were larger than when the noise variables were excluded. This pattern was present, though less marked, when the interaction was simple and was only slightly apparent in the case of no interaction.

Correlation between predicted and actual values for cross-validation sample

Table 4 includes the correlation between the actual and predicted values of the outcome variable in the cross-validation sample by the method of prediction, level of model complexity, sample size and the strength of the relationship. Perhaps the most notable pattern here is that the correlations for GAM and MARS are uniformly above 0.95, regardless of the simulation condition. The correlations for NNET, while not as high as those for GAM and MARS, were generally above 0.90 across conditions, and above 0.95 in all situations where model complexity was not at its highest level. On the other hand, the correlations between the actual and predicted values for CART and BOOST were never as high as 0.95, with those of CART not even attaining 0.91 for any condition simulated here. In the case of BOOST the correlation between the actual and predicted values were lower than for any other method with the complex interaction model, were comparable to those of CART in the middle model complexity condition, and somewhat higher than CART for the least complex model. In no situation simulated here were the BOOST correlations comparable to those of the other studied methods, however. For both OLS and CORR, the correlations between the predicted and actual values were comparable to one another across simulation conditions. In addition, for the least complex model, the correlation values for OLS and CORR were comparable to those from GAM, MARS and NNET, and somewhat lower than these three methods in the simple interaction case.

Table 4. *Correlation between Actual and Cross-Validation Values*

Interaction	N	Strength	OLS	CART	GAM	MARS	COR	BOOST	NNET
Complex	50	.2	0.872	0.759	0.986	0.953	0.843	0.615	0.895
		.5	0.874	0.762	0.992	0.959	0.845	0.611	0.901
		.8	0.875	0.761	0.992	0.959	0.846	0.620	0.899
	100	.2	0.878	0.828	0.992	0.978	0.846	0.622	0.944
		.5	0.881	0.828	0.996	0.982	0.848	0.629	0.939
		.8	0.881	0.827	0.996	0.982	0.848	0.623	0.933
	500	.2	0.883	0.893	0.996	0.992	0.845	0.631	0.973
		.5	0.886	0.898	0.998	0.995	0.846	0.634	0.969
		.8	0.886	0.897	0.998	0.995	0.848	0.639	0.964
	1000	.2	0.884	0.905	0.995	0.993	0.846	0.635	0.975
		.5	0.887	0.905	0.998	0.996	0.848	0.635	0.976
		.8	0.888	0.905	0.998	0.996	0.850	0.636	0.978
Simple	50	.2	0.920	0.832	0.978	0.968	0.927	0.859	0.952

		.5	0.930	0.840	0.989	0.984	0.935	0.868	0.978
		.8	0.930	0.840	0.990	0.979	0.936	0.868	0.973
	100	.2	0.924	0.868	0.981	0.978	0.927	0.865	0.977
		.5	0.932	0.876	0.992	0.990	0.935	0.874	0.993
		.8	0.934	0.876	0.993	0.991	0.937	0.876	0.995
	500	.2	0.927	0.895	0.985	0.984	0.927	0.871	0.987
		.5	0.937	0.904	0.994	0.993	0.936	0.880	0.997
		.8	0.935	0.904	0.991	0.995	0.935	0.879	0.998
	1000	.2	0.928	0.898	0.985	0.985	0.927	0.870	0.988
		.5	0.936	0.906	0.994	0.993	0.936	0.879	0.997
		.8	0.935	0.908	0.995	0.995	0.937	0.882	0.998
None	50	.2	0.988	0.840	0.985	0.979	0.988	0.927	0.974
		.5	0.998	0.849	0.998	0.997	0.997	0.938	0.996
	100	.8	0.999	0.850	0.999	0.999	0.999	0.939	0.999
		.2	0.988	0.867	0.987	0.986	0.988	0.928	0.983
		.5	0.998	0.878	0.998	0.998	0.997	0.938	0.997
	500	.8	0.999	0.879	0.999	0.999	0.998	0.940	0.999
		.2	0.988	0.888	0.988	0.988	0.988	0.929	0.988
		.5	0.998	0.897	0.998	0.998	0.997	0.939	0.998
	1000	.8	0.999	0.897	0.999	0.999	0.998	0.940	0.999
		.2	0.988	0.891	0.988	0.988	0.988	0.929	0.988
		.5	0.998	0.900	0.998	0.998	0.997	0.939	0.998
		.8	0.999	0.902	0.999	0.999	0.999	0.940	0.999

The results in Table 5 show the correlations between the actual and predicted values of the response variable by model complexity, level of noise and strength of the relationship between the predictors and the outcome. Correlations obtained from most of the methods studied here, including OLS, CART, GAM, MARS and CORR were not influenced by the presence of noise variables in the model. On the other hand, the correlations between observed and predicted values from the BOOST method were lower in the presence of noise than in its absence. This pattern was particularly notable for the most complex model. The predicted with actual correlations produced by NNET were negatively influenced by the presence of noise in the model for the most complex model but not under for the other model conditions.

Table 5. *Correlation between Actual and Predicted Values by Prediction Method, Level of Interaction, Level of Noise in the Model and Strength of Relationship*

Interaction	Noise	strength	OLS	CART	GAM	MARS	COR	BOOST	NNET
Complex	no	.2	0.881	0.847	0.992	0.978	0.845	0.818	0.968
		.5	0.884	0.851	0.996	0.983	0.847	0.820	0.967
		.8	0.884	0.849	0.996	0.983	0.847	0.823	0.964
	yes	.2	0.877	0.845	0.994	0.980	0.845	0.433	0.925
		.5	0.879	0.845	0.997	0.983	0.847	0.434	0.926
		.8	0.881	0.846	0.997	0.983	0.848	0.436	0.923
Simple	no	.2	0.926	0.874	0.983	0.978	0.927	0.903	0.983
		.5	0.935	0.882	0.992	0.990	0.936	0.912	0.996
		.8	0.935	0.883	0.993	0.993	0.936	0.913	0.997

	yes	.2	0.924	0.872	0.983	0.979	0.927	0.830	0.968
		.5	0.932	0.881	0.992	0.990	0.935	0.838	0.987
		.8	0.932	0.881	0.993	0.987	0.936	0.839	0.985
None	no	.2	0.988	0.873	0.987	0.985	0.988	0.966	0.985
		.5	0.998	0.882	0.998	0.998	0.997	0.976	0.997
		.8	0.999	0.882	0.999	0.999	0.998	0.977	0.999
	yes	.2	0.988	0.870	0.987	0.985	0.988	0.891	0.981
		.5	0.998	0.880	0.998	0.998	0.997	0.901	0.997
		.8	0.999	0.881	0.999	0.999	0.998	0.902	0.999

Discussion

The current study is among the first to systematically compare the performance of such a comprehensive set of the relatively new prediction models using Monte Carlo simulated conditions. In prior research, single datasets were commonly used, and only a small number of these alternative prediction methods were examined simultaneously. In the current study, we have attempted to include a much larger set of these methods and have compared them all under a common set of known conditions using the Monte Carlo approach. The goal of this work is to help inform researchers about optimal methods for obtaining predicted values in a variety of conditions. The results presented here have several implications for researchers interested in obtaining predicted values for continuous response variables. First, it is clear that the level of model complexity has an impact on the performance of several of the methods studied here. For example, when the population model for the response consists only of a linear combination of the independent variables with no interactions, CORR and BOOST produced the lowest RMSE values for the cross-validated data, followed closely by OLS, GAM, and MARS. In addition, in this least complex case predictions obtained from CORR and OLS had slightly higher correlations with the actual cross-validation data than did the other methods, although correlations for both GAM and MARS were only marginally lower. It would appear, therefore, that if the researcher knows with some certainty that no interactions are present in the population, then using CORR for prediction in concert with standardized values of the independent and dependent variables might be a successful strategy for developing a prediction model. This positive result for the CORR method is in keeping with results reported in Dana and Dawes (2004). It should be emphasized that while when taken together these results for the simplest case do support CORR as producing optimal prediction results, several other methods, including BOOST, OLS, GAM and MARS provided predicted values that were almost as accurate as CORR. Nonetheless, none of these other approaches are as simple as using the correlation coefficients, so that parsimony of the model may support this technique as compared to the others.

When the population model contained an interaction of two independent variables and/or a squared or cubed term, the performance of both CORR and BOOST was degraded substantially. In the simple interaction condition, OLS continued to perform very well when compared with the other methods studied here, despite the fact that the model used did not actually contain the interaction. In this case, the OLS predictions produced the lowest RMSE for the cross-validated data values consistently except for the largest sample size, in which case GAM had a comparable RMSE. The correlations between the predicted and actual values were higher than OLS for some of the more complex prediction algorithms, such as GAM, MARS and NNET, however. In contrast to the simple interaction model, when a more complex relationship existed between the independent and dependent variables, GAM, and to a lesser extent MARS, displayed a clear advantage in terms of their ability to accurately predict the response for the cross-validated data. Indeed, in this most complex case, GAM consistently produced the lowest RMSE and highest correlation values regardless of sample size, strength of relationship between predictors and response and the presence or absence of noise variables. These results support earlier findings in the physical sciences for GAM (Moisen & Frescino, 2002) and MARS (Munoz & Felicisimo, 2004).

The relatively poor performance of CART was somewhat surprising given positive results reported in earlier studies (Lee & Jin, 2006). However, it should be noted that much of this earlier research involved

actual, rather than simulated data. Furthermore, in these earlier studies the researchers engaged in pruning of the tree prior to coming up with the final predictive model. Pruning, which involves the removal of nodes that are deemed statistically unnecessary, is an important aspect of tree building because it helps to prevent overfitting of the training data (Breiman, Friedman, Olshen, & Stone, 1984). Such overfitting can lead to a tree model that can very accurately predict values for the training sample but that is not generalizable to other samples from the population. Indeed, this appears to be what occurred in the current study. A brief examination of the ability of CART to predict response values for the training samples in this study revealed that its RMSE was much lower than the RMSE for the cross-validation samples. Therefore, one conclusion that should be drawn from these results is that researchers using CART must become adept at pruning trees in order to produce one that is sufficiently generalizable to the broader population.

In summary, then, it would appear that a researcher who uses the standard OLS methodology for most prediction problems will obtain reasonably accurate values unless the underlying model is fairly complex. On the other hand, one could always rely on a more sophisticated approach, such as GAM and MARS, and be assured that regardless of the underlying population model the resulting predictions will be optimal or very close to optimal in most cases. Of course, there are practical limitations to using these more complex methods, including a lack of familiarity on the part of many applied researchers. In addition, the greater sophistication of some of these approaches may require greater knowledge of how they work so that appropriate settings are used. This would appear to be particularly true for CART, and to a lesser extent NNET. This increased difficulty in use may make the more complex techniques somewhat less appealing than the more familiar and simple OLS and CORR methods, particularly when it is known or strongly suspected that the underlying model is not complex.

Limitations and directions for future research

There are limitations to the current study that should be addressed in future research. The simulated data came from the normal distribution, so it is not known how well these various methods would perform when the assumption of normality has been violated. In addition, both the response and predictor variables were continuous in nature. Future studies should focus on the ability of these methods to accurately predict group membership for a categorical outcome, and their ability to use categorical predictor variables as well. As noted above, some of the methods such as CART, require greater researcher intervention to be successfully employed. Future studies should attempt to incorporate the pruning process into the simulation algorithm, although we recognize that automating this could present substantial logistical challenges. Finally, it is important to note that the sole goal in this study was to optimize prediction accuracy, with no attention paid to model parsimony. While in practice some applications of prediction models share this goal, in other instances researchers wish to balance such accuracy concerns with model parsimony. In these cases, a researcher may be interested in finding the smallest possible set of independent variables that yields very accurate predictions (though perhaps not optimal). When this combination of parsimony and accuracy is of interest, researchers may want to consider methods such as best subsets regression, or variants thereof.

References

Berk, R. A. (2008). *Statistical learning from a regression perspective*. New York: Springer.

Breiman, L., Friedman, J. H., Olshen, R. A., & Stone, C. J. (1984). *Classification and regression trees*. Monterey, CA: Wadsworth & Brooks/Cole Advanced Books & Software.

Buhlmann, P., & Hothorn, T. (2007). Boosting algorithms: Regularization, prediction and model fitting. *Statistical Science, 22*, 477-505.

Buhlmann, P., & Yu, B. (2003). Boosting with the L_2 loss: Regression and classification. *Journal of the American Statistical Association, 98*, 324-339.

Dana, J., & Dawes, R. M. (2004). The superiority of simple alternatives to regression for social science predictions. *Journal of Educational and Behavioral Statistics, 29*, 317-331.

Dawes, R. M. (1979). The robust beauty in improper linear models in decision making. *American Psychologist, 34*, 571-582.

Dawes, R. M., & Corrigan, B. (1974). Linear models in decision making. *Psychological Bulletin, 81,* 95-106.

Dempster, A. P., Schatzoff, M., & Wermuth, N. (1977). A simulation study of alternatives to ordinary least squares. *Journal of the American Statistical Association, 72,* 77-91.

Finch, H., & Schneider, M. K. (2007). Classification accuracy of neural networks vs. discriminant analysis, logistic regression and classification and regression trees: Three and five groups cases. *Methodology, 3,* 47-57.

Freund, Y., & Schapire, R. (1997). A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of Computer System Science, 55,* 119-139.

Garson, G. D. (1998). *Neural networks: An introductory guide for social scientists*. London: SAGE Publications.

Goldberg, L. R. (1972). Parameters of personality inventory construction and utilization: A comparison of prediction strategies and tactics. *Multivariate Behavioral Research Monographs*. Fort Worth, TX: Society of Multivariate Experimental Psychology.

Hastie, T., Tibshirani, R., & Friedman, J. (2001). *The elements of statistical learning: Data mining, inference and prediction*. New York: Springer.

Hill, R. W., & Holland, P. W. (1977). Two robust alternatives to least-squares regression. *Journal of the American Statistical Association, 72,* 828-833.

Hussain, S. S., & Sprent, P. (1983). Non-parametric Regression. *Journal of the Royal Statistical Society A, 146,* 182-191.

Kuhnert, P. M., Mengersen, K., & Tesar, P. (2003). Bridging the gap between different statistical approaches: An integrated framework for modeling. *International Statistical Review, 71,* 335-368.

Lee, S., & Jin, S. (2006). Decision tree approaches for zero-inflated count data. *Journal of Applied Statistics, 33,* 853-865.

Marks, M. R. (1966, September). *Two kinds of regression weights which are better than betas in crossed samples*. Paper presented at the meeting of the American Psychological Association, New York.

Marshall, D. B., & English, D. J. (2000). Neural network modeling of risk assessment in child protective services. *Psychological Methods, 5,* 102-124.

Moisen, G. G., & Frescino, T. S. (2002). Comparing five modeling techniques for predicting forest characteristics. *Ecological Modeling, 157,* 209-225.

Munoz, J., & Felicisimo A. M. (2004). Comparison of statistical methods commonly used in predictive modeling. *Journal of Vegetation Science, 15,* 285-292.

Nguyen, N., & Cripps, A. (2001). Predicting house value: A comparison of multiple regression analysis and artificial neural networks. *Journal of Real Estate Research, 22,* 313-336.

Park, B. U., & Kim, W. C. (1997). Simple transformation techniques for improved non-parametric regression. *Scandinavian Journal of Statistics, 24,* 145-163.

R Development Core Team (2007). *R: A language and environment for statistical computing.* Vienna: R Foundation for Statistical Computing.

Razi, M. A., & Athappily, K. (2005). A comparative predictive analysis of neural networks (NNs), nonlinear regression and classification and regression tree (CART) models. *Expert Systems with Applications, 29,* 65-74.

Reibnegger, G., Weiss, G., Werner-Felmayer, G., Judmaier, G., & Wachter, H. (1991). Neural networks as a tool for utilizing laboratory information: Comparision with linear discriminant analysis and with classification and regression trees. *Proceedings of the National Academy of Science, 88,* 11426-11430.

Schumacher, M., Robner, R., & Vach, W. (1996). Neural networks and logistic regression: Part I. *Computational Statistics and Data Analysis, 21,* 661-682.

Smith, A. E., & Mason, A. K. (1997). Cost estimation predictive modeling: Regression versus neural network. *The Engineering Economist, 42,* 137-160.

Tukey, J. (1977). *Exploratory data analysis.* Reading, MA: Addison-Wesley.

Waller, N. G., & Jones, J. A. (2009). Correlation weights in multiple regression. *Psychometrika,* Retrieved from http://www.springerlink.com/content/e3572021626270x6/

Williams, C. J., Lee, S. S., Fisher, R. A., & Dickerman, L.H. (1999). A comparison of statistical methods for prenatal screening for down syndrome. *Applied Stochastic Models in Business and Industry*, 15, 89-101.

Wood, S. N. (2006). *Generalized additive models: An introduction with R.* Boca Raton, FL: Chapman & Hall/CR.

Zientek, L. R., & Thompson, B. (2009). Matrix summaries improve research reports: Secondary analyses using published literature. *Educational Researcher, 38,* 343-352.

Send correspondence to: Holmes Finch or Jocelyn Holden
Ball State University
Email: whfinch@bsu.edu or jeholden@bsu.edu

APPENDIX

Example R Code for Simulations

```
library(MASS)
library(tree)
library(nnet)
library(mgcv)
library(mboost)
library(mda)
wais3 <- matrix(
c(1, .76, .58, .43,
.76, 1, .57, .36,
.58, .57, 1, .45,
.43, .36, .45, 1),
nrow=4,ncol=4)
NumSubj<-500
NumSubj2<-500
set.seed(125810470)
finalresults.rmse.simulation11<-NULL
finalresults.cor.simulation11<-NULL
for(z in 1:1000) {
wais.data <- mvrnorm(n = NumSubj, mu=rep(0,14), Sigma=wais3, empirical = TRUE)
EigVecs <- eigen(wais3)$vectors
EigVals <- eigen(wais3)$values
PC.Weights.1 <- matrix(EigVecs[, 1] * 1/sqrt(EigVals[1]), 14,1)
PC.Weights.14 <- matrix(EigVecs[,14] * 1/sqrt(EigVals[14]),14,1)
PC.Scores.1 <- scale(wais.data %*% PC.Weights.1)
PC.Scores.14 <- scale(wais.data %*% PC.Weights.14)
#yhat.1 is collinear with the last principal component
wais.data.frame<-data.frame(wais.data)
attach(wais.data.frame)
y1 <- .2*X1 + .2*X2+ .2*X3 + .2*X4 + rnorm(NumSubj2,mean=0,sd=.1)
wais.cross <- mvrnorm(n = NumSubj2, mu=rep(0,14), Sigma=wais3, empirical = TRUE)
EigVecs <- eigen(wais3)$vectors
EigVals <- eigen(wais3)$values
PC.Weights.1 <- matrix(EigVecs[, 1] * 1/sqrt(EigVals[1]), 14,1)
PC.Weights.14 <- matrix(EigVecs[,14] * 1/sqrt(EigVals[14]),14,1)
PC.Scores.1 <- scale(wais.cross %*% PC.Weights.1)
PC.Scores.14 <- scale(wais.cross %*% PC.Weights.14)
#yhat.1 is collinear with the last principal component
wais.cross.frame<-data.frame(wais.cross)
wais.cross.boost.frame<-data.frame(wais.cross.frame$X1, wais.cross.frame$X2,
wais.cross.frame$X3, wais.cross.frame$X4)
wais.cross.frame.boost<-
data.frame(wais.cross.frame$X1,wais.cross.frame$X2,wais.cross.frame$X3,wais.cross.frame$X4)
y1.cross <- .2*wais.cross.frame$X1 + .2*wais.cross.frame$X2 + .2*wais.cross.frame$X3 +
.2*wais.cross.frame$X4 + rnorm(NumSubj2,mean=0,sd=.1)
waiscross.predict.ols<-predict(lm(y1~X1+X2+X3+X4),wais.cross.frame)
waiscross.predict.tree<-predict(tree(y1~X1+X2+X3+X4),wais.cross.frame)
waiscross.predict.gam<-predict(gam(y1~s(X1)+s(X2)+s(X3)+s(X4)),wais.cross.frame)
wais.mars<-mars(wais.data, y1)
waiscross.predict.mars<-predict(wais.mars,wais.cross)
```

```
waiscross.predict.nnet2<-predict(nnet(y1~X1+X2+X3+X4,size=2,linout=T,
skip=T),wais.cross.frame)
waiscross.predict.nnet3<-predict(nnet(y1~X1+X2+X3+X4,size=3,linout=T,
skip=T),wais.cross.frame)
waiscross.predict.nnet4<-predict(nnet(y1~X1+X2+X3+X4,size=4,linout=T,
skip=T),wais.cross.frame)
waiscross.predict.nnet5<-predict(nnet(y1~X1+X2+X3+X4,size=5,linout=T,
skip=T),wais.cross.frame)
waiscross.predict.nnet15<-predict(nnet(y1~X1+X2+X3+X4,size=15,linout=T,
skip=T),wais.cross.frame)
rxy <- cor(y1, wais.data)
waiscross.predict.cor<-
wais.cross.frame$X1*rxy[1,1]+wais.cross.frame$X2*rxy[1,2]+wais.cross.frame$X3*rxy[1,3]+w
ais.cross.frame$X4*rxy[1,4]
y1.boost<-as.numeric(y1)
waiscross.boost<-glmboost(y1.boost~X1+X2+X3+X4, control=boost_control(mstop=1000,
center=FALSE))
waiscross.predict.boost<-
coefficients(waiscross.boost)[2]*wais.cross.frame$X2+coefficients(waiscross.boost)[3]*wais.cro
ss.frame$X3+coefficients(waiscross.boost)[4]*wais.cross.frame$X4
y1.cross2<-as.numeric(y1.cross)
predicted2.values<-data.frame(y1.cross, waiscross.predict.ols, waiscross.predict.tree,
waiscross.predict.cor, waiscross.predict.gam, waiscross.predict.mars,
waiscross.predict.nnet2,waiscross.predict.nnet3, waiscross.predict.nnet4,
waiscross.predict.nnet5, waiscross.predict.nnet15, waiscross.predict.boost)
attach(predicted2.values)
ols.rmse<-sqrt((1/50)*sum(waiscross.predict.ols-y1.cross2)^2)
cart.rmse<-sqrt((1/50)*sum(waiscross.predict.tree-y1.cross2)^2)
gam.rmse<-sqrt((1/50)*sum(waiscross.predict.gam-y1.cross2)^2)
mars.rmse<-sqrt((1/50)*sum(waiscross.predict.mars-y1.cross2)^2)
cor.rmse<-sqrt((1/50)*sum(waiscross.predict.cor-y1.cross2)^2)
boost.rmse<-sqrt((1/50)*sum(waiscross.predict.boost-y1.cross2)^2)
nnet2.rmse<-sqrt((1/50)*sum(waiscross.predict.nnet2-y1.cross2)^2)
nnet3.rmse<-sqrt((1/50)*sum(waiscross.predict.nnet3-y1.cross2)^2)
nnet4.rmse<-sqrt((1/50)*sum(waiscross.predict.nnet4-y1.cross2)^2)
nnet5.rmse<-sqrt((1/50)*sum(waiscross.predict.nnet5-y1.cross2)^2)
nnet15.rmse<-sqrt((1/50)*sum(waiscross.predict.nnet15-y1.cross2)^2)
results.rmse<-
data.frame(ols.rmse,cart.rmse,gam.rmse,mars.rmse,cor.rmse,boost.rmse,nnet2.rmse,nnet3.rmse,n
net4.rmse,nnet5.rmse,nnet15.rmse)
finalresults.rmse.simulation11<-rbind(finalresults.rmse.simulation11,results.rmse)

cor.simulation<-cor(y1.cross, predicted2.values)
cor.simulation<-data.frame(cor.simulation)
finalresults.cor.simulation11<-rbind(finalresults.cor.simulation11,cor.simulation)
}
mean(finalresults.rmse.simulation11)
mean(finalresults.cor.simulation11)
```

Canonical Correlation Analysis: A Step-by-Step Example in Commonly Available Software

Eric L. Oslund
Texas A & M University

Canonical Correlation Analysis (CCA) can be conceptualized as a multivariate regression involving multiple outcome variables. CCA compares two sets of variables and is the second-most general application of the General Linear Model (GLM) following Structural Equation Modeling. Structural Equation Modeling software have made conducting CCA feasible for researchers in numerous and disparate disciplines. SPSS and AMOS are two commonly used statistical software packages and both can conduct CCAs. AMOS is particularly useful because it enables statistical significance testing of individual canonical correlations. This article provides background knowledge of CCA as part of the GLM and provides step-by-step instructions for conducting a CCA in AMOS.

The general linear model (GLM) comprises an overwhelming majority of parametric statistical procedures. Regression is the univariate GLM (Cohen, 1968) and canonical correlation analysis (CCA) is the multivariate GLM (Knapp, 1978). The GLM includes the *t*-test, ANOVA and other OVA procedures, descriptive discriminate analysis, CCA and structural equation modeling (SEM; Graham, 2008; Henson, 2002; Thompson, 1991; Zientek & Thompson, 2009). SEM is the acme of the hierarchical GLM, subsuming all other procedures in the GLM. CCA is the second-most general application of the GLM (Henson, 2002; Knapp, 1978; Sherry & Henson, 2005; Thompson, 1991). All GLM procedures are defined by the fact that they a) create weights applied to measured variables to construct synthetic variables, b) are correlational and c) provide analogues of the r^2 effect size (Henson, 2002; Kellow, 1998; Thompson, 2006).

Although univariate methods can be done in both CCA and SEM due to the hierarchical nature of the GLM, typically CCA and SEM focus on multivariate analyses. Multivariate procedures are crucial to the behavioral sciences for several reasons. Arguably the biggest reason is that multivariate methods honor the complex world of behavioral sciences research (Fan, 1997; Sherry & Henson, 2005; Thompson, 1991). Human behavior is complex and has multiple causes and multiple finalities. Multivariate statistical procedures can aid in bridging the gap between the theoretical and practical world of behavioral sciences. Multivariate methods provide valuable information that univariate methods simply cannot (Thompson, 1985).

Another major reason for use of multivariate procedures is that it keeps Type I error to a minimum (Fan, 1997; Sherry & Henson, 2005; Thompson, 1985, 1991). Experimentwise Type I error rate grows considerably the more statistical tests are conducted on the same dataset (as an example, running multiple post hoc test in ANOVA inflates the experimentwise error rate unless a correction procedure is used to adjust testwise alpha levels). Multivariate methods such as CCA and SEM protect against this inflation.

There are three main purposes of this article. The first is to provide the reader with a brief background on CCA. Thompson (1985) pointed out that CCA is becoming more and more popular as an analysis tool in different disciplines. Illustrating that CCA can be done in SEM will also help the reader understand the connection of multivariate methods in the GLM.

The second purpose is to provide the reader with detailed directions on how to use the AMOS 5.0 software, which is a SEM program, to conduct a CCA. It should be noted that AMOS 18.0 yields the same results as AMOS 5.0; however, AMOS 5.0 was used as it is currently available for free. Krus, Reynolds, and Krus (1976) observed that CCA became more popular through the use of computer programs; this is even truer today than it was 30 years ago. Brief instructions on use of SPSS and SAS will be given because using the three programs concurrently to compare results is recommended. Also, there are several outputs that one program gives and the others do not. Sherry and Henson (2005) discuss the importance of using step-by-step instructions to further the use of statistical analyses.

The final purpose is to show the reader how using AMOS 5.0 allows for testing of individual canonical correlations, canonical function coefficients, canonical structure coefficients, and index coefficients through use of the bootstrap procedure. Testing individual canonical correlations is an attractive feature available in AMOS 5.0 and is not done in SPSS.

CCA Overview

CCA employs the Multiple Indicators/Multiple Causes (MIMIC) model, which postulates that indicators are both caused by (effect indicators), and cause (causal indicators) synthetic variables (MacCallum & Browne, 1993). The MIMIC model is essential to CCA and the creation of the synthetic variates that compose canonical functions. In CCA, sets of measured variables are compared by creating synthetic variables though linear combinations of the measured variables (Fan, 1997; Graham, 2008; Henson, 2002; Knapp, 1978; Sherry & Henson, 2005). Through the synthetic variables, the relationship between sets of measured variables is examined.

The creation of the synthetic variables is analogous to computing synthetic $\hat{Y}$ scores in regression (Henson, 2002). As throughout the GLM, weights are applied to the measured variables to calculate the synthetic variables (called canonical variables in CCA). However, unlike regression procedures, the linear-combined synthetic canonical variables (or variates) are calculated without error (Fan, 1997).

The first canonical function is composed of two synthetic variables, one from the predictor variables and one from the dependent variables (Graham, 2008). The number of canonical functions is equal to the number of measured variables in the smaller set (Graham, 2008; Thompson, 1991). The Pearson *r* correlation, called the canonical correlation, between the two synthetic variables in a canonical function is the primary focus of CCA (Sherry & Henson, 2005). The square of the canonical correlation represents the amount of variance shared by the variables that compose the canonical correlation (Thompson, 1984) and is analogous to the R^2 effect size from multiple regression (Sherry & Henson, 2005). This reiterates the GLM.

The construction of the synthetic variables is done in a way to maximize the canonical correlations (Fan, 1997; Graham, 2008; Knapp, 1978; Sherry & Henson, 2005; Thompson, 1984). In order to do this, the maximum likelihood estimation method is used as opposed to the least squares estimation method (which maximizes explained variance) (Graham, 2008). These canonical functions are similar to components from principal component analysis (Thompson, 1991) and factors from exploratory factor analysis (Henson, 2002).

After the first canonical correlation is produced, there remains residual variance in the variable sets that is unexplained by the first canonical function. All subsequent canonical functions are composed to analyze the residual variance left over from the preceding canonical function (Graham, 2008; Henson, 2002; Sherry & Henson, 2005). Of course, if the first canonical correlation explains 100% of the variance (r = 1.0), the researcher can stop there. The synthetic variables comprised in a canonical function are orthogonal (uncorrelated) to all other synthetic variables outside their own canonical function (Graham, 2008; Sherry & Henson, 2005; Thompson, 1984, 1991). That each canonical function is orthogonal to both predictor and dependent synthetic variables of all other canonical functions is a concept called *double orthogonality* (Sherry & Henson, 2005).

In the GLM, weights are applied to measured variables to produce the synthetic variable(s). In CCA, these weights are called canonical function coefficients, either standardized or unstandardized. These weights are analogous to beta weights in regression. They are contextual and vary based on other measured variables in the model and multicollinearity (Sherry & Henson, 2005; Thompson, 2006). Due to their being contextual and because CCA is a multistep process, the functional coefficients are constrained throughout the analyses. Index coefficients, or cross loadings, are the correlation of measured variables with the synthetic variable of the opposite set (Fan, 1997; Graham, 2008; Thompson, 1984).

Structure coefficients, which are the correlations between a measured variables and the synthetic variable, are crucial to interpretation of CCA results (Dunlap & Landis, 1998; Graham, 2008; Sherry & Henson, 2005; Thompson, 1984, 1991). They are a measure of the direct effect a measured variable has on a synthetic variable and should always be interpreted because of the relativity of function coefficients. As Henson (2002) articulated, "if standardized weights inform the researcher what variables are getting credit for the effect, then structure coefficients inform the researcher what variables could have gotten credit for the effect" (p. 11). Only interpreting function coefficients may lead to falsely believing one variable is more, or less, important than another when that may not be the case.

A historic problem with conducting CCA analysis through most statistical software is that there was no way to test for the statistical significance of individual canonical correlations or the various coefficients (Bagozzi, Fornell, & Larcker, 1981; Sherry & Henson, 2005; Thompson, 1991). For

example, if there are three canonical functions, only the last test of statistical significance is for a single canonical correlation (for canonical function three). The first test of statistical significance test all functions together; the second test function two and three combined. Only when the last function is statistically significant are all the previous functions also statistically significant (Sherry & Henson, 2005).

One very distinct advantage of conducting and CCA in SEM is that statistical significance testing can be done for individual canonical correlations and also for function, structure, and index coefficients (Fan, 1997; Graham, 2008). According to Knapp (1978):

> The principle objective of canonical correlation analysis is to find a linear combination of the *p* variables that correlates maximally with a linear combination of the *q* variables and, for sample data, to test the statistical significance of that correlation. (p. 411)

Others may argue that statistical significance testing is not as important as other indications of meaningful results, such as effect sizes. However, that SEM allows for testing statistical significance is a significant advantage. In order conduct the various statistical significance tests, bootstrap procedures must be employed.

Bootstrap Overview

Bootstrapping allows for resampling of data for both descriptive and inferential use (Fan & Wang, 1996; Thompson, 1993). The original dataset is copied, as many times as the user specifies, and then added to the dataset. This process creates a huge dataset from which resamples are drawn (with replacement) of exactly the same size as the original sample (Thompson, 1995, 2006). In inferential applications, the number of resamples drawn should number at least 2,000 and larger resamples are preferable (Fan, 2003; Thompson, 1995, 1999); samples of this size are easily done by AMOS 5.0.

Bootstrapping procedures create an empirically-based sampling distribution when theoretical distributions do not exist (Fan & Wang, 1996; Thompson, 1999). A sampling distribution created from a bootstrap does not make, nor require, as many assumptions about the data as a theoretical distribution (Thompson, 1999). An empirically-based sampling distribution is important if those assumptions are violated (Fan & Wang, 1996). In the case of CCA, bootstrapping is absolutely necessary because theoretical distributions are nonexistent for function, structure, and index coefficients (Fan & Wang, 1996). From this empirical sampling distribution, standard errors are calculated, which play several important roles in CCA.

First, the standard errors provide information about the precision of observed estimates (Dunlap & Landis, 1998; Fan, 1997; Thompson, 1999). Dunlap and Landis (1998) state, "the interpretation of weight significance depends on knowledge of the standard error of the weight" (p. 398). Two equal function coefficients of .50 would have wildly different interpretations if one coefficient's standard error was .90 and the other's was .001.

Standard errors also allow for testing of statistical significance (Bagozzi, et al., 1981; Graham, 2008; Thompson, 1999). This is simply done by dividing the observed parameter estimate (a function coefficient, canonical correlation, etc.) by the standard error of the estimate, which produces the critical ratio (Bagozzi, et al., 1981; Thompson, 2006). An estimate with a critical ratio equal to approximately 1.96, based on the standard normal distribution ($M = 0$, $SD = 1$), is statistically significant in a two-tailed test against a nil-null hypothesis. AMOS 5.0, as an SEM program, easily performs this function while SPSS and SAS do not.

Step-by-Step CCA Example Done in AMOS

The following is an example demonstrating how to run a CCA analysis using the free student version of AMOS 5.0. The Holzinger and Swineford (1939) data were used and can be conveniently found at (http://www.coe.tamu.edu/~bthompson/datasets.htm#1). The dependent variables used were Math Number Puzzles (hereafter T21) and Math Word Problem Reasoning (hereafter T22). The predictor variables were General Information Verbal Test, Paragraph Comprehension Test, and Sentence Completion Test (hereafter T5, T6 and T7 respectively). There is no theoretical basis for comparing these variable sets; in real applications, there should always be thought given to the basis for comparing variable sets through CCA (Sherry & Henson, 2005). The reader is encouraged to replicate the current example and to always check results in two different statistical packages as the computations are difficult and lend themselves to easy mistakes being made.

The analysis was also done in SPSS and SAS for comparison purposes; the syntax for SPSS and SAS provided in the Appendix. CCA can also be run in SAS through a series of point-and-click selections. After opening SAS, choose the "Solutions" drop-down menu, click on the "Analysis" option and then the "Analyst" option. Open the data file then go to the "Statistics" drop down menu, choose "Multivariate" and then "Canonical Correlation". In the box that opens up, choose variables T5, T6, and T7 for "Set 1" and T21 and T22 for "Set 2". Click "OK" to run the analysis.

The data was in an SPSS file and imported into AMOS using the "Data Files" option in the "Files" drop-down menu. The variables were then examined for selection using the "Variables in Dataset" option under the "View/Set" drop-down menu. This option brings up every variable in the dataset from which the relevant variables can be imported and analyzed. All following procedures in this example will utilize the visual icons located on the furthest left pane in AMOS. The drop-down menu tabs at the top of the AMOS frame can also be utilized for those more comfortable with that method.

The figure for the first synthetic variable for the predictor variables, labeled LF1 (indicating it is the left side of the first canonical function) is drawn by clicking and dragging the measured variables from the "Variables in Dataset" box. Click on the variable of interest in the "Variables in Dataset" box and drag it into the modeling plane and release it. This is done for the remainder of measured variables. Figure 1 illustrates the complete drawing for the first left synthetic variable using the MIMIC model.

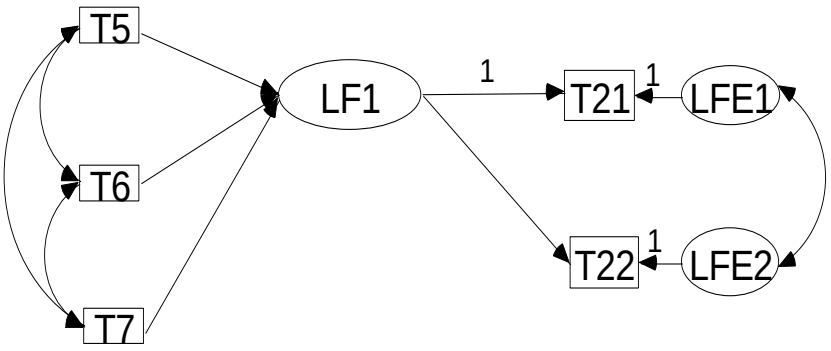

Figure 1. CCA diagram representing the first synthetic variable from the predictor variable set.

The synthetic variable is then created and named. The user selects the "draw unobserved variables" icon, which is represented by an oval in the top-left portion of the AMOS window. This variable must be named at the user's discretion; however, indicating in some manner that it is a synthetic variable for the first canonical function is recommended for clarity. As more variables are added, the figures become progressively busier and labeling can be important to allow the user and others to easily identify variables.

Dependent variables must have a synthetic variables created and attached to represent the residual/error variance. This is done by selecting the "add a unique variable to an existing variable" icon, which is represented by a box with an attached circle above it on the top-left portion of the screen. Move the figure over the dependent variables and click one time for each dependent variable. The default path value from these latent variables to their respective measured variables will be 1.0. Once these variables are created, they need to be named. In the current example, the error variances are labeled LFE1 and LFE2 for T21 and T22 respectively. LFE is short for left-function error. Again, these labels are given at the discretion of the researcher.

The paths between variables must be drawn and a marker variable must be determined in order to run the analysis. Select the directional path represented by the icon of a line with a single point, which AMOS calls "draw paths (single headed arrows)". Single-headed arrows must be drawn from the predictor variables (T5, T6, and T7) to LF1 and from LF1 to the dependent variables T21 and T22. Choose one of the paths from LF1 to a dependent variable (in this case, T21) and set the value to 1.0. This is accomplished by right-clicking on the path from LF1 to T21, clicking on "parameters" tab, and setting the "regression weight" to 1.0. LF1 is now in the same metric as T21 and considered the marker variable.

The covariance's between the variables, represented by non-directional (double arrow) paths are then drawn between every possible combination of T5, T6, and T7. This is repeated for LFE1 and LFE2. The analysis properties must then be established once the drawing is complete.

Click on the "analysis properties" icon located in the middle column and roughly midway down the page. The default estimation method is "maximum likelihood" (and was used in this example); however,

this can be changed on the "estimation" tab. Under the "output" tab, check the boxes for the "standardized estimates" and "all implied moments". Close the window and then click the "calculate estimates" icon located immediately to the right of the "analysis properties" icon. An error message is likely to appear indicating that the endogenous synthetic variable (in this case, LF1) does not have an error variable attached. This refers to a variable similar to the error variables LFE1 and LFE2 for criterion variables T21 and T22. Click on the "proceed with the analysis" option. This effectively sets the error variance of LF1 at zero, which can also be accomplished by creating a residual/error variance latent variable and constraining the variance to zero. Simply ignoring the error message is easier.

Once the analysis is run, click the "view text" icon. This gives the output for the analysis. A "decimals" drop-down menu is on the top output file; the setting should be set so that the output is to the 12th decimal. This is critical as minor rounding error can cause inaccurate outputs in subsequent analysis. Table 1 provides the unstandardized canonical function coefficients (called unstandardized regression weights in AMOS 5.0) of the individual paths including the standard errors, critical ratios and associated *p* values using the *z* distribution. In this example, all individual regression paths are statistically significant except the path from LF1 to T21. This is because the path value was constrained and not freely estimated.

Table 1. Statistical Significance Tests of Unstandardized Canonical Function Coefficients (UCFC).

UCFC	Path estimate	Standard error	Critical ratio	*P* value
T5 to LF1	.037752036695	.018214782033	2.072604361989	.038209114766
T6 to LF1	.180258686233	.067915221761	2.654172092166	.007950326870
T7 to LF1	.127449262517	.049433912843	2.578174681851	.009932377397
LF1 to T21	1.0	-	-	-
LF1 to T22	2.926949942591	.478352643185	6.118812102930	<.001

Note. In the row LF1 to T21, - indicate values that were not freely estimated due to the path estimate being constrained to 1.0 as the marker variable.

The next step of the analysis requires the same procedure be followed for the synthetic variable of the second canonical function (called LF2 in this example) with the exception that the directional paths for LF1 are constrained to the observed values from the previous analysis (provided in Table 1). In addition to the paths being constrained to the unstandardized function weights, a marker variable must be set. This follows the aforementioned procedure of constraining the regression weight for the path from LF2 to a criterion variable to 1.0; in this example the path from LF2 to T21 is constrained. Figure 2 represents the drawing, including constraints, for the two left synthetic variables.

The path from T5 to LF1 is constrained using the observed value (from Table 1) to the 12th decimal place, which equals .037752036695. The remaining constrained paths in the Figure 3 display the path estimate values to the 2nd decimal place. The values should never be constrained to the second decimal place and are done so here for the sole purpose of being able to easily see which paths are being constrained. Always use the observed values to the 12th decimal place as shown in the T5-LF1 path.

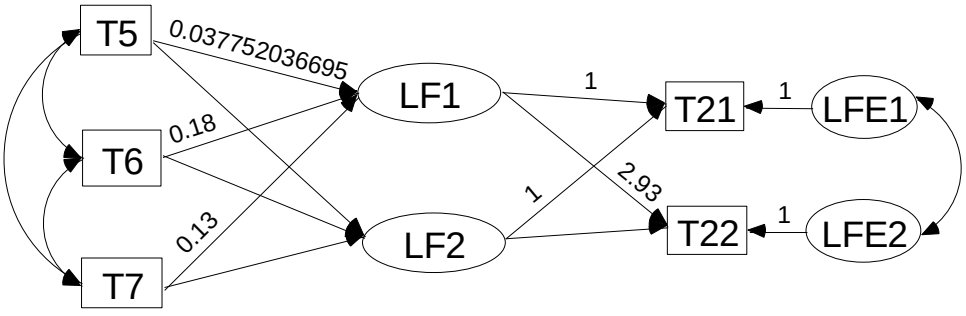

Figure 2. CCA diagram for second synthetic variable from the predictor variables set. The paths are constrained to the values obtained from the analysis of the first synthetic variable of the predictor variable set.

The next step is to draw the figure for the first right synthetic variable (RF1). This is represented in Figure 3 and is a duplicate of Figure 1 with the exception that the directional paths are reversed. Essentially, the regression diagram is reversed and the criterion variables (T21 and T22) from the previous analyses now become the exogenous variables and the predictor variables (T5, T6, and T7) become the criterion variables. It should be noted SPSS still labels the initial criterion variables as "dependent".

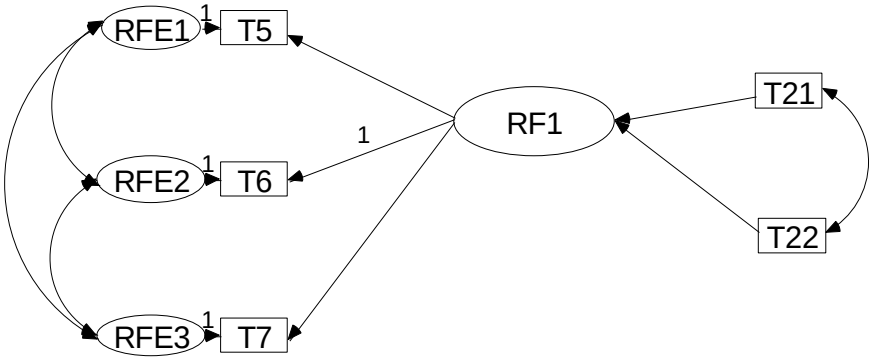

Figure 3. CA diagram representing the first synthetic variable from the dependent variable.

The same analyses are done for RF1 and RF2 as were done for LF1 and LF2. Once RF1 canonical function coefficients (unstandardized regression weights) are obtained, the associated paths are constrained in the RF2 analysis as demonstrated in Figure 4. The paths are shown here to the 2nd decimal for ease of viewing; however, again it is crucial to use values to the 12th decimal and to also set a marker variable for each separate analysis.

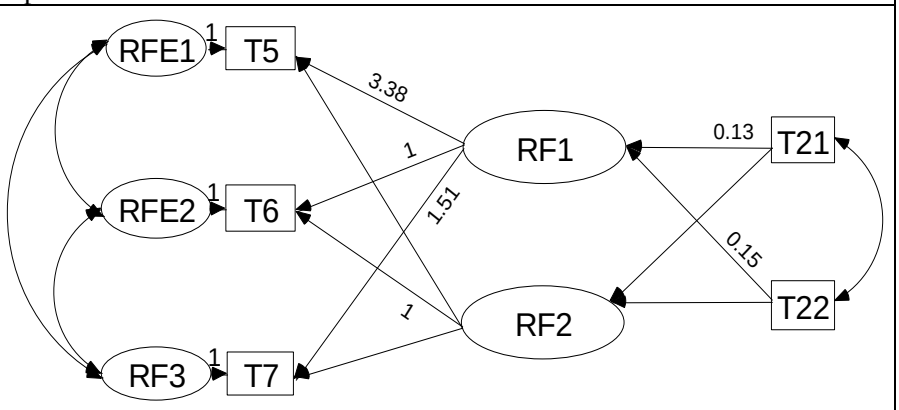

Figure 4. CCA diagram for second synthetic variable from the dependent variable set. The paths are constrained to the values obtained from the analysis of the first synthetic variable from the dependent variable set.

The next step requires combining the completed diagrams from the LF2 and RF2 analyses as demonstrated in Figure 5. The diagram includes regression paths from T5-7 to the left synthetic variables and from T21-22 to the right synthetic variables. The regression paths are constrained to the values from the analyses conducted previously. The canonical function coefficients for the left variables are presented to the 2nd decimal place for ease of viewing only.

The next step involves correlating all the measured variables as depicted in Figure 6. The constraints/estimates for the function coefficient paths are not shown so the reader can easily see all the paths. The paths would be constrained to the observed

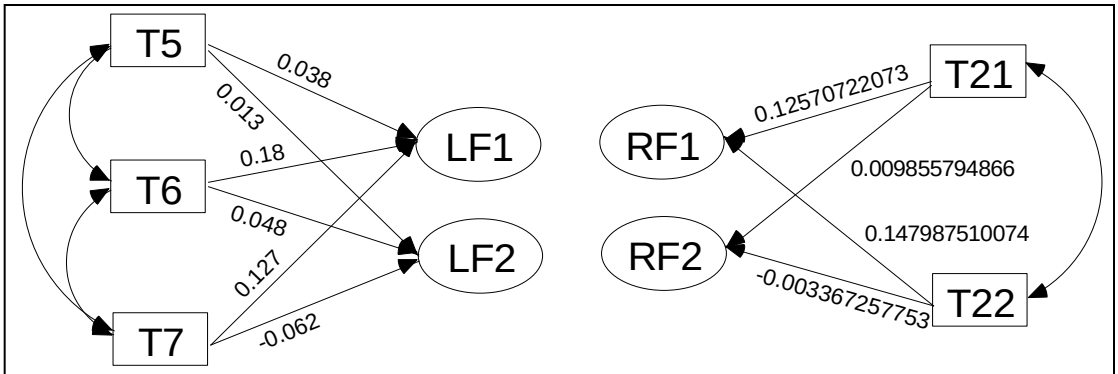

Figure 5. CCA diagram with the canonical function coefficient paths constrained to previously observed values.

values. The canonical correlations (correlations between LF1 and RF1 and between LF2 and RF2) cannot be drawn because AMOS does not allow drawing correlations between synthetic variables. The analysis can be run at this point.

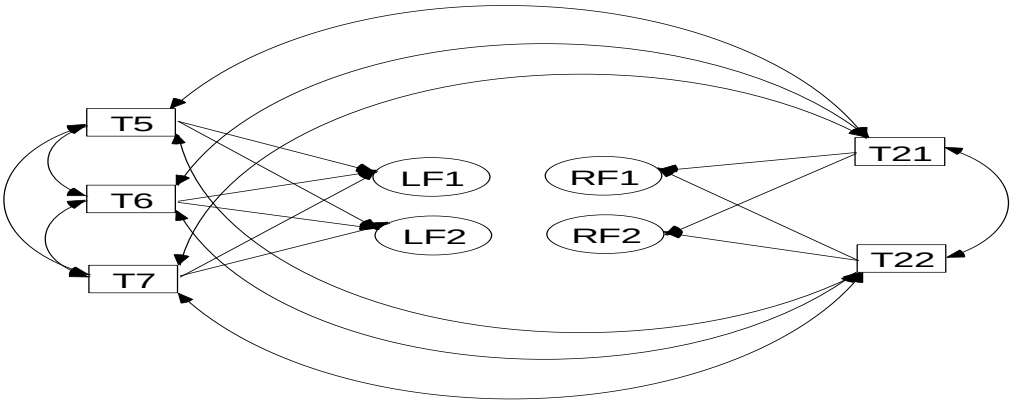

Figure 6. Final CCA diagram including all paths. The canonical function coefficient path values are omitted to ease viewing of diagram. The paths would be constrained using the values found in Figure 5.

On the "analysis properties" icon, select the "bootstrap" tab. Check the "perform bootstrap" and "bias-corrected confidence intervals" boxes. It is recommended to run at least 2,000 bootstrap samples and use a 95% bias corrected (BC) con-fidence interval. Bias-corrected estimation is preferred because it makes fewer as-sumptions regarding the sampling distribution (Thompson, 1993). Under the "output" tab, the standardized estimates (stand-ardized beta weights or stand-ardized canonical function coefficients) and "all implied moments" boxes should be checked.

Interpretation.

The bootstrap procedure allows for an empirically derived sampling distribution to be calculated, which allows for statistical significance testing, building of confidence intervals, and calculating standard errors. This is a particular appealing feature in AMOS that SPSS does not readily do. Arguably the most important difference between the two programs is that AMOS allows for statistical significance testing of individual canonical correlations as opposed to testing sequential correlations as done by SPSS.

The output file provides numerous analyses values. Of particular interest are the standardized beta weights (a.k.a., standardized canonical function coefficients), canonical structure coefficients, and canonical correlations. Standard errors, confidence intervals, and *p* values for the statistical significance tests for these statistics are also provided in disparate tables.

Table 2 provides the bivariate correlations, structure coefficients, index coefficients and canonical correlations for the completed CCA taken from the "Implied (for all variables) Correlations" table. Index coefficients will not be discussed here other than to mention they are the bivariate correlations between the measured variables of one set with the synthetic variables of another.

Table 2. Bivariate Correlations, Structure Coefficients, Index Coefficients and Canonical Correlations

	T22	T21	T7	T6	T5	RF2	RF1	LF2	LF1
T22	-								
T21	0.377[a]	-							
T7	0.470[a]	0.302[a]	-						
T6	0.448[a]	0.321[a]	0.733[a]	-					
T5	0.426[a]	0.307[a]	0.716[a]	0.657[a]	-				
RF2	-0.319[b]	0.758[b]	-0.022[c]	0.013[c]	0.014[c]	-			
RF1	0.948[b]	0.653[b]	0.488[c]	0.477[c]	0.454[c]	0.00[e]	-		
LF2	-0.018[c]	0.044[c]	-0.383[b]	0.219[b]	0.244[b]	0.058[d]	0.00[e]	-	
LF1	0.501[c]	0.345[c]	0.923[b]	0.901[b]	0.858[b]	0.00[e]	0.529[d]	0.00[e]	-

Note. [a]Bivariate correlations. [b]Structure coefficients. [c]Index coefficients. [d]Canonical correlations. [e]Reflect double orthogonality.

Notice the correlation of LF1 with RF2 is 0.00, as is the correlation between LF1 with LF2. This demonstrates double orthogonality and reflects the rule of CCA that the correlations between all synthetic variables of the first canonical function are completely uncorrelated with all synthetic variables in the second function and all subsequent functions (Sherry & Henson, 2005).

The first canonical function, LF1 with RF1, is equal to .529. The second function between LF2 and RF2 is equal to .058. While it is impossible to discern statistical significance at a glance, it is immediately apparent that the first function is much more likely to be statistically significant that the second against the null hypothesis that $r = 0.00$. Comparing the canonical correlations given by SPSS, AMOS, and SAS is recommended to ensure the analysis was done correctly. If the canonical correlations match between the two programs, it can be assumed that the analysis was done correctly and that the remaining AMOS output, such as the tests of significance for paths and canonical correlations, is accurate. In addition to the correlation, SPSS and SAS also give the squared correlation value under "Eigenvalues and canonical correlations" and "Canonical Correlation Analysis" respectively.

A critical step in any CCA analysis is the examination of structure coefficients as they are a measurement of the direct impact of a measured variable on a synthetic variable. Both AMOS and SPSS provide the structure coefficients, which again should be compared to ensure accuracy. In Table 3, the correlation between T22 and function one ($r = .948$) is a structure coefficient. SPSS provides the same output, calling them the correlations between either the "dependent" or "covariates" and the "canonical variables".

Table 3. Standardized Canonical Coefficients and Structure Coefficients

	Std Canonical Coefficients		Structure Coefficients	
Variable	Function 1	Function 2	Function 1	Function 2
T21	.345	1.023	.653	.758
T22	.818	-.705	.948	-.319
T5	.297	.836	.858	.244
T6	.400	.841	.901	.219
T7	.418	-1.598	.923	-.383

Comparing the standardized canonical function coefficients with the structure coefficients for function one found in Table 3, it is apparent that the difference between the two on T5 is quite large (.297 vs. .858). Does this mean T5 is a much better predictor, based on the structure coefficient equaling .858, than it appears to be if only the standardized function coefficient of .297 where interpreted? Because the proportion of T5 to T6 to T7 is roughly the same on both the structure and the function coefficients, and because the ranking of the values from least to greatest is the same, interpreting both the structure and function coefficients would lead to the same conclusion. The best predictor for function one is T7 and the worse is T5.

Statistical Significance Testing. In the final analysis output, AMOS provides the *p* values for the significance tests (against H_o: $r = 0.00$) for the bivariate correlations, index coefficients, structure coefficients, and canonical correlations. The tests for individual function coefficients are conducted along the previous steps of the analysis (see Table 1 for an example). The ability to test these statistics in AMOS makes it appealing as SPSS does not readily compute these tests of significance. The AMOS output also comprises standard error estimates and the lower- and upper-bound estimates for confidence intervals.

Table 4 replicates the AMOS table "Implied (for all variables) Correlations – Two Tailed Significance (BC)", providing the *p* values for statistical significance testing. The first canonical function (correlation between LF1 and RF1) is statistically significant as $p = .001$. This was somewhat

Table 4.Observed p Values Testing H_0: $r = 0.00$

	T22	T21	T7	T6	T5	RF2	RF1	LF2
T21	.001							
T7	.001	.001						
T6	.001	.001	.001					
T5	.001	.001	.001	.001				
RF2	.001	.001	.719	.813	.761			
RF1	.001	.001	.001	.001	.001	.978		
LF2	.772	.437	.001	.001	.001	.275	.983	
LF1	.001	.001	.001	.002	.001	.992	.001	.998

expected due to their correlation of .529. The second canonical function (r = .058), fails to reach statistical significance (p = .275). SPSS does not test individual canonical correlations, except for the last one. The only time to know the statistical significance of all individual canonical correlations using SPSS is when the last canonical correlation is statistically significant. If the last one is statistically significant, then all individual canonical correlations are significant.

The p values for the statistical significance tests between LF1 and LF2 and between LF1 and RF2 are .998 and .992 (see Table 4). According to the double orthogonality rule, these values should be 1.0. The presence of sampling error prevents the probability of rejecting the null from being 0.00% (Graham, 2008), which is the theoretical value. This fact further highlights the need to constrain paths throughout the model to the highest decimal place available to reduce the amount of sampling error contamination.

The structure and index coefficients also have reported p values as seen in Table 4. All the structure coefficients are statistically significant with the highest p value being .002 for T6 with LF1. Exactly half of the index coefficients are statistically significant. As the number of bootstrap procedures increases, the p values decrease as a function of smaller standard errors and larger critical ratios.

The caveat for statistical significance testing throughout the entire GLM applies in CCA; statistical significance is directly related to sample size and, in many ways, test whether you have a numerous sample (Thompson, 1991). SEM typically requires large sample sizes, which likely inflates the significance of statistically significant findings. However, because many journals are reluctant to publish results that fail to reach statistical significance, this article provides the heuristic demonstration of how to test statistical significance for individual canonical correlations, structure coefficients and function coefficients. As Max Planck (1949) once said, "A new scientific truth does not triumph by convincing its opponents and making them see the light, but rather because its opponents eventually die, and a new generation grows up that is familiar with it" (p. 33-34). Perhaps this will be true of statistical significance testing across the GLM.

Caution is also needed when interpreting effect sizes. The r^2 effect size for the first canonical correlation is .280. If holding rigidly to conventional standards, most would conclude this is a small effect size. However, effect sizes must be considered in light of previous research findings and thoughtfulness when determining if the magnitude of effect is meaningful. If the effect size is three times larger than any other previous research has found, suddenly an effect size of .280 becomes very meaningful and quiet large. Likewise, if the observed effect size means the difference in extending the life of a cancer patient by five years, it is extremely meaningful. As with all statistical analyses, it is at least as important and probably more to accurately interpret results then it is to be able to run analyses.

Discussion

CCA is a useful technique for comparing two sets of variables and is part of the GLM. Recent SEM computer software and computer capabilities have made it easier to conduct CCAs. The purpose of this paper was to expand the reader's knowledge of CCA and give explicit instructions and a replicable example.

A few important considerations are mentioned to reiterate key principles expounded upon in this paper. Canonical function coefficients and structural coefficients should always be examined together as only interpreting function coefficients can lead to erroneous conclusions. As an SEM software package, AMOS is user-friendly and allows for the statistical significance testing of individual canonical correlations, function, structure, and index coefficients, which other statistical software does not readily do. It is recommended, especially the first few times a CCA is conducted, to check the results across multiple software packages.

The uses of statistical techniques take time to be adopted following their invention. As Cohen (1992) pointed out, the t-test took nearly 40 years to be adopted and included in statistical textbooks. One way to quicken the use of statistical techniques, in this case CCA, is to provide the background knowledge and perhaps more importantly, detailed instructions and examples.

Explicit directions and examples are especially critical in the evolving world of increasingly sophisticated software that allows statistical procedures to be done in times inconceivable even a few years ago. However, knowledge of using a software program must be buttressed with knowledge of the analysis being run. There is no point, and much danger, in running an analysis that isn't understood. The reader will hopefully not only replicate the examples provided in this paper, but also read other

informative articles and books (many referenced in this paper) to gain both the knowledge of what a CCA is and how to conduct one.

References

Bagozzi, R. P., Fornell, C., & Larcker, D. F. (1981). Canonical correlation analysis as a special case of a structural relations model. *Multivariate Behavioral Research, 16*, 437-454. doi:10.1207/s15327906mbr1604_2

Cohen, J. (1968). Multiple regression as a general data-analytic system. *Psychological Bulletin, 70*, 426-433. doi:10.1037/h0026714

Cohen, J. (1992). A power primer. *Psychological Bulletin, 112*, 155-159. doi:10.1037/0033-2909.112.1.155

Dunlap, W. P., & Landis, R. S. (1998). Interpretations of multiple regression borrowed from factor analysis and canonical correlation. *Journal of General Psychology, 125*, 397-407. doi:10.1080/00221309809595345

Fan, X. (1997). Canonical correlation analysis and structural equation modeling: What do they have in common? *Structural Equation Modeling, 4*, 65-79. doi:10.1080/1070551970954 0060

Fan, X. (2003). Using commonly available software for bootstrapping in both substantive and measurement analyses. *Educational and Psychological Measurement, 63*, 24-50. doi:10.1177/0013164402239315

Fan, X., & Wang, L. (1996). Comparability of jackknife and bootstrap results: An investigation for a case of canonical correlation analysis. *Journal of Experimental Education, 64*, 173-190.

Graham, J. M. (2008). The general linear model as structural equation modeling. *Journal of Educational and Behavioral Statistics, 33*, 485-506. doi:10.3102/1076998607306151

Henson, R. K. (2002, April). *The logic and interpretation of structure coefficients in multivariate general linear model analyses*. Paper presented at the Annual Meeting of the American Educational Research Association, New Orleans.

Holzinger, K. J., & Swineford, F. (1939). *A study in factor analysis: The stability of a bi-factor solution* (Vol. 48): Chicago: University of Chicago. doi:10.1086/440440

Kellow, J. T. (1998). Beyond statistical significant tests: The importance of using other estimates of treatment. *American Journal of Evaluation, 19*, 123-135. doi:10.1177/1098214098019 00112

Knapp, T. R. (1978). Canonical correlation analysis: A general parametric significance-testing system. *Psychological Bulletin, 85*, 410-416. doi:10.1037/0033-2909.85.2.410

Krus, D. J., Reynolds, T. J., & Krus, P. H. (1976). Rotation in canonical variate analysis. *Educational and Psychological Measurement, 36*, 725-730. doi:10.1177/0013164476036 00320

MacCallum, R. C., & Browne, M. W. (1993). The use of causal indicators in covariance structure models: Some practical issues. *Psychological Bulletin, 114*, 533-541. doi:10. 1037/0033-2909.114.3.533

Planck, M. (1949). *Scientific autobiography and other papers*. New York: Philosophical Library. doi: 10.1016/S0016-0032(49)90297-5

Sherry, A., & Henson, R. K. (2005). Conducting and interpreting canonical correlation analysis in personality research: A user-friendly primer. *Journal of Personality Assessment, 84*, 37-48. doi:10.1207/s15327752jpa8401_09

Thompson, B. (1984). *Canonical correlation analysis: Uses and interpretation*. Newbury Park, CA: Sage. doi:10.1016/0191-8869(87)90162-0

Thompson, B. (1985, April). *Heuristics for teaching multivariate general linear model techniques*. Paper presented at the Annual Meeting of the American Educational Research Association, Chicago.

Thompson, B. (1991). A primer on the logic and use of canonical correlation analysis. *Measurement and Evaluation in Counseling and Development, 24*, 80-95.

Thompson, B. (1993). The use of statistical significance tests in research: Bootstrap and other alternatives. *Journal of Experimental Education, 61*, 361-377.

Thompson, B. (1995). Exploring the replicability of a study's results: Bootstrap statistics for the multivariate case. *Educational and Psychological Measurement, 55*, 84-94. doi:10.1177 /0013164495055001008

Thompson, B. (1999, April). *Common methodology mistakes in educational research revisited: Along with a primer on both effect sizes and bootstrap*. Paper presented at the Annual meeting of the American Educational Research Association, Montreal.

Thompson, B. (2006). *Foundations of behavioral statistics: An insight-based approach*. New York: Guilford Press.

Zientek, L. R., & Thompson, B. (2009). Matrix summaries improve research reports: Secondary analyses using published literature. *Educational Researcher, 38,* 343-352. doi:10. 3102/0013189X09339056

Zientek, L. R., & Thompson, B. (2009). Matrix summaries improve research reports: Secondary analyses using published literature. *Educational Researcher, 38,* 343-352.

Send correspondence to: Eric L. Oslund
Texas A & M University
Email: ericoslund@yahoo.com

APPENDIX

SPSS and SAS Syntax for the Current CCA Example

SPSS syntax

```
manova t21 t22 with t5 t6 t7
/PRINT=SIGNIF(MULTIV UNIV EIGEN DIMENR)
  /DISCRIM(STAN COR ALPHA(.999))
  /method=unique
  /design .
execute.
```

SAS syntax

```
proc cancorr data='C:\HolzingerData.sas7bdat';
var t5 t6 t7;
with t21 t22;
run; quit;
```

All Possible Kappa Coefficient Values and Cell Distribution Combinations in a 2 x 2 Matrix: The Case of the Small Sample

David A. Walker
Northern Illinois University

This research provided practitioners and researchers with information containing all possible kappa coefficient values and cell distribution combinations for 2 x 2 matrices comprised of binary data and with a sample size of N = 5. It was found that there was a low probability of 0.16 when employing a small sample size that a kappa value would reside in the categories that signified $\kappa \geq 0.60$ from the Landis and Koch (1977) or the Fleiss (1971) and Cicchetti (1984) scales, with a probability ratio indicating that kappa values were 5.25 times more likely to be derived from the bottom categories of these scales where $\kappa < 0.60$.

Inter-rater agreement indices used in 2 x 2 classification matrices with categorical data have been studied comprehensively in the social science scholarly literature (Berry & Mielke, 1988; Brennan & Prediger, 1981; Fleiss, 1975; Light, 1971; Soeken & Prescott, 1986; Zwick, 1988). A sampling from the literature in the field of education shows that inter-rater agreement has been applied in classroom-based contexts pertaining to in-class student behaviors (Junod, DuPaul, Jitendra, Volpe, & Cleary, 2006; Nolan, Gadow, & Sverd, 1994), to examine social communication (Olswang, Svensson, Coggins, Beilinson, & Donaldson, 2006), or to look at teaching and classroom processes (Walker de Felix, Waxman, Paige, & Huang, 1993).

There are numerous indices for inter-rater agreement that can be applied to 2 x 2 tables with categorical data. Cohen's (1960) kappa coefficient, or κ, is cited frequently as an index used in inter-rater agreement studies (Umesh, Peterson, & Sauber, 1989; Zwick, 1988). Traditionally, kappa has been employed as a chance-corrected index related to the proportion of inter-rater agreement when using a 2 x 2 matrix with two raters and binary data. In Figure 1, the main diagonal (i.e., cells A and D) indicates the agreement level between the raters as either 00 or 11 and the off diagonal (i.e., cells B and C) indicates the level of disagreement between the raters as either 10 or 01.

		Rater 2		
		0	1	**Total**
Rater 1	0	Cell A	Cell B	$\mathbf{p_1}$
	1	Cell C	Cell D	$\mathbf{q_1}$
	Total	$\mathbf{p_2}$	$\mathbf{q_2}$	**n**

Assumptions affiliated with κ are that "N objects categorized are independent; the assigners operate independently; and the categories are independent, mutually exclusive, and exhaustive" (Brennan & Prediger, 1981, p. 688). Mathematically, kappa can be derived from the following formula (Subkoviak, 1988):

$$\kappa = \frac{P_O - P_E}{1 - P_E} \quad (1)$$

where, P_O = observed agreement, where $P_O = (A + D)/N$; A = count from cell 1; D = count from cell 4; N = number of observations; B = count from cell 2; C = count from cell 3; and P_E = expected percentage of agreement based on chance $= [(A + B)*(A + C) + (C + D)*(B + D)]/N^2$.

The maximum value of kappa is 1.00, which indicates total agreement between two raters. The minimum value for kappa is 0.0, which signifies that there is not better than chance that the two raters would agree. Although kappa can take on negative values to -1.00, which indicate agreement worse than expected by chance, said negative values tend to be reported in the literature as 0.0, thus, keeping kappa at a functional range from 0.0 to 1.00. Landis and Koch (1977) provide a guide concerning general interpretations of kappa values, where inter-rater agreement values are categorized as:

	< 0.0		= Poor
0.0	to	0.20	= Slight
0.21	to	0.40	= Fair
0.41	to	0.60	= Moderate
0.61	to	0.80	= Substantial
0.81	to	1.00	= Almost Perfect

Other researchers, such as Fleiss (1971) and Cicchetti (1984), have presented similar guidelines for interpreting inter-rater agreement values derived from kappa-like coefficients:

	< 0.40	=	Poor
0.40	to 0.59	=	Fair
0.60	to 0.74	=	Good
0.75	to 1.00	=	Excellent

However, as with guidelines provided for effect size measures (*cf*. Cohen, 1988; Glass, McGaw, & Smith, 1981), the context of the research and the use of the results in a clinical or academic setting should be evaluated before decisions are made pertaining to the magnitude of a particular kappa value.
Cohen (1960) noted that the kappa coefficient can reach a maximum value of 1.00, or perfect agreement between two raters, only when cells B and C have zero counts in them. That is, when the degree of rater disagreement is zero. Along with Cohen's work, Brennan and Prediger (1981) and Umesh et al. (1989) examined, although not extensively, some of the maximum values that kappa could achieve based on the distributions of inter-rater agreement and disagreement in cells A through D in a 2 x 2 matrix with binary data. What these studies emphasized was the fact that in many instances, such as when marginals are free or as Brennan and Prediger found "… a margin is 'free' whenever the marginal proportions are not known to the assigner beforehand" (p. 690), there are sundry possible values for a kappa coefficient, which is problematic in interpreting agreement results. As Umesh et al. (1989) noted about the importance of understanding kappa coefficient values and cell distribution derived from a study situation, "… in general, knowing the maximum value of a kappa statistic facilitates inferences about its practical or substantive significance" (p. 845).

Purpose and Methods

Adding to the research provided by Cohen (1960), Brennan and Prediger (1981), and Umesh et al. (1989), a purpose of this study was to provide practitioners and researchers with detailed information containing all possible kappa coefficient values and cell distribution combinations for a 2 x 2 matrix comprised of binary data, free marginals, and an observation size of N = 5. A small observational sample was a primary application because it represented recurrent sizes in applied research fields (Huberty & Mourad, 1980), especially in classroom observation situations where small samples often are encountered.

Further, this study used a small observation size example due to the enormity of possible ratings in a 2 x 2 matrix with binary data, two raters, and free marginals. That is, the case of N = 5 has 2^{10} or 1,024 possible ratings, N = 10 has 1,048,576 ratings or 2^{20}, N = 15 has 1,073,741,824 ratings or 2^{30}, and N = 20 has 1,099,511,627,776 ratings or 2^{40}. For the case of N = 5, the various cell combinations were determined by letting the rows represent the ratings for rater 1 and the columns represent the ratings for rater 2 as noted in Figure 1. To determine the possible number of combinations within a matrix, for example, if we had an arrangement in cells A through D of 3002 (i.e., both raters had three 0s and two 1s or $\kappa = 1.00$), we would have 10 possible combinations for this cell distribution: 11100, 11010, 11001, 10101, 10110, 10011, 01011, 01101, 01110, 00111. Statistical Analysis System code provided by Klar, Lipsitz, Parzen, and Leong (2002) and Mundfrom (D. Mundfrom, personal communication, April 29, 2008) were used for this research intent.

A second purpose for this research was to study the use of the Landis and Koch (1977) and Fleiss (1971)/Cicchetti (1984) interpretive guides as a comparative condition pertaining to the degree of inter-rater agreement when determining the probability of attaining kappa values ≥ 0.60 for the case of N = 5. A kappa value ≥ 0.60 has been noted in the scholarly literature, both theoretically (Cicchetti; Fleiss, 1971; Landis & Koch; Subkoviak, 1988) and in application (Lavigne et al., 1994; Schrijnemaekers & Haverman, 1993), as a threshold between "moderate/good" inter-rater magnitude of agreement and the converse of this agreement extent.

Results and Discussion

With five observations, the sample space for all negative, zero, and positive kappa values was 56. Table 1 shows the entire sample space when N = 5. To keep congruent with the scholarly literature in reporting kappa values that range from 0.0 to 1.00, the sample space was re-operationalized as only non-negative kappa values and became 38. Of the 38 possible kappa coefficient values, for both the Landis and Koch (1977) and Fleiss (1971)/Cicchetti (1984) parameters, 6 were ≥ 0.75, with all achieving $\kappa =$

1.00 or perfect agreement, and 2 were ≥ 0.60 with κ = .62 in both instances. From Table 1, we can see that the previously-noted 8 instances where κ ≥ .60 had 92 possible combinations that resulted in this desired cut-off or the numerator. Of the possible 1,024 ratings, 574 returned non-negative κ values or the denominator. Thus, in the case of N = 5, the probability of obtaining a kappa value ≥ 0.60 on either the Landis and Koch or the Fleiss-Cicchetti scales was only 0.16.

When all possible combinations of cell distributions and kappa values are known, Table 1 shows that there was a greater chance of obtaining kappa values in the bottom ranges of rater agreement, where κ < 0.60. There was a high probability when employing a small sample size, that a kappa value would reside in either the "moderate," "fair," or "slight" categories from the Landis and Koch (1977) scale or the "fair" or "poor" categories from the Fleiss (1971) and Cicchetti (1984) scales (i.e., all < 0.60). The probability ratio (i.e., P(κ < 0.60) or 0.840 / P(κ ≥ 0.60) or .160) indicated this result by showing that with 5 observations from the Landis and Koch or the Fleiss/Cicchetti scales, kappa values were 5.25 times more likely to be derived from the bottom categories where κ < 0.60.

Therefore, when using either of these guidelines to describe the relative strength of agreement for kappa coefficient values, researchers and practitioners using a small set of observations should not be surprised if the value of κ is determined not to have great magnitude of agreement due to the lower probability of achieving a desired level of agreement ≥ 0.60. That is, using adjectives from either guideline of "substantial," "good," "almost perfect," or "excellent" to describe the magnitude of agreement between two raters at a defined threshold of κ ≥ 0.60 may not be appropriate when using kappa with small sets of observations. This recommendation can be extended beyond the small data set example of N = 5 because when sample size increases, the ratio gap of N to sample space is augmented due to exponentially higher possible ratings for 2^{20} (N = 10), 2^{30} (N = 15) or, 2^{40} (N = 20) and power is lost, which leads to a lower probability that κ ≥ 0.60.

Table 1. *All Cell Distributions and Possible Combinations for N = 5*

Kappa	SE	P_O	P_E	Cell A	Cell B	Cell C	Cell D	Possible Combinations
1.00*	0.0	1.00	1.00	5	0	0	0	1
1.00*				0	0	0	5	1
1.00*	.45	1.00	.52	3	0	0	2	10
1.00*				2	0	0	3	10
1.00*	.45	1.00	.68	4	0	0	1	5
1.00*				1	0	0	4	5
0.62*	.41	.80	.48	2	0	1	2	30
0.62*				2	1	0	2	30
0.55	.40	.80	.56	1	0	1	3	20
0.55				1	1	0	3	20
0.55				3	0	1	1	30
0.55				3	1	0	1	20
0.29	.31	.60	.44	1	0	2	2	25
0.29				2	2	0	1	25
0.29				1	2	0	2	30
0.29				2	0	2	1	30
0.17	.45	.60	.52	1	1	1	2	60
0.17				2	1	1	1	60
0.12	.21	.40	.32	1	3	0	1	20
0.12				1	0	3	1	20
0	0.0	.80	.80	0	1	0	4	5
0				4	1	0	0	5
0				4	0	1	0	5
0				0	0	1	4	5
0	0.0	.60	.60	0	0	2	3	10
0				3	2	0	0	10
0				3	0	2	0	10

Table 1 (continued). *All Cell Distributions and Possible Combinations for N = 5*

Kappa	SE	P_O	P_E	Cell A	Cell B	Cell C	Cell D	Possible Combinations
0				0	2	0	3	10
0	0.0	.40	.40	0	3	0	2	10
0				2	3	0	0	10
0				0	0	3	2	10
0				2	0	3	0	10
0	0.0	.20	.20	0	0	4	1	5
0				1	4	0	0	5
0				0	4	0	1	5
0				1	0	4	0	5
0	0.0	0.0	0.0	0	0	5	0	1
0				0	5	0	0	1
-0.15	.41	.40	.48	1	2	1	1	60
-0.15				1	1	2	1	60
-0.25	.45	.60	.68	3	1	1	0	20
-0.25				0	1	1	3	20
-0.36	.40	.40	.56	0	2	1	2	30
-0.36				2	2	1	0	30
-0.36				2	1	2	0	20
-0.36				0	1	2	2	30
-0.43	.31	.20	.44	1	1	3	0	20
-0.43				0	3	1	1	20
-0.43				1	3	1	0	25
-0.43				0	1	3	1	25
-0.47	.21	0.0	.32	0	4	1	0	5
-0.47				0	1	4	0	5
-0.67	.45	.20	.52	0	2	2	1	30
-0.67				1	2	2	0	30
-0.92	.41	0.0	.48	0	3	2	0	10
-0.92				0	2	3	0	10

Note: SE = standard error derived from Fleiss' (1981) corrected standard error, P_O = observed agreement, P_E = expected percentage of chance agreement. * = $\kappa \geq 0.60$

References

Berry, K. J., & Mielke, P. W. (1988). A generalization of Cohen's kappa agreement measure to interval measurement and multiple raters. *Educational and Psychological Measurement, 48*, 921-933.

Brennan, R. L., & Prediger, D. J. (1981). Coefficient kappa: Some uses, misuses, and alternatives. *Educational and Psychological Measurement, 41*, 687-699.

Cicchetti, D. V. (1984). On a model for assessing the security of infantile attachment: Issues of observer reliability and validity. *Behavioral and Brain Sciences, 7*, 149-150.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement, 20*, 37-46.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Lawrence Erlbaum Associates, Publishers.

Fleiss, J. L. (1971). Measuring nominal scale agreement among many raters. *Psychological Bulletin, 76*, 378-382.

Fleiss, J. L. (1975). Measuring agreement between two judges on the presence or absence of a trait. *Biometrics, 31*, 651-659.

Fleiss, J. L. (1981). *Statistical methods for rates and proportions* (2nd ed.). New York: Wiley & Sons.

Glass, G. V., McGaw, B., & Smith, M. L. (1981). *Meta-analysis in social research*. Beverly Hills, CA: Sage.

Huberty, C. J., & Mourad, S. A. (1980). Estimation in multiple correlation/prediction. *Educational and Psychological Measurement, 40*, 101-112.

Junod, R. E. V., DuPaul, G. J., Jitendra, A. K., Volpe, R. J., & Cleary, K. S. (2006). Classroom observations of students with and without ADHD: Differences across types of engagement. *Journal of School Psychology, 44*, 87-104.

Klar, N., Lipsitz, S. R., Parzen, M., & Leong, T. (2002). An exact bootstrap confidence interval for κ in small samples. *The Statistician, 51*, 467-478.

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics, 33*, 159-174.

Lavigne, J. V., Arend, R., Rosenbaum, D., Sinacore, J., Cicchetti, C., Binns, H. J., Christoffel, K. K., Hayford, J. R., & McGuire, P. (1994). Interrater reliability of the DSM-III-R with preschool children. *Journal of Abnormal Child Psychology, 22*, 679-690.

Light, R. J. (1971). Measures of response agreement for qualitative data: Some generalizations and alternatives. *Psychological Bulletin, 76*, 365-377.

Nolan, E. E., Gadow, K. D., & Sverd, J. (1994). Observations and ratings of tics in school settings. *Journal of Abnormal Child Psychology, 5*, 579-593.

Olswang, L. B., Svensson, L., Coggins, T. E., Beilinson, J. S., & Donaldson, A. L. (2006). Reliability issues and solutions for coding social communication performance in classroom settings. *Journal of Speech, Language, and Hearing Research, 49*, 1058-1071.

Schrijnemaekers, M. A., & Haverman, M. J. (1993). Depression in frail Dutch elderly: The reliability of the Zung scale. *Clinical Gerontologist, 13*, 59-66.

Soeken, K. L., & Prescott, P. A. (1986). Issues in the use of kappa to estimate reliability. *Medical Care, 24*, 733-741.

Subkoviak, M. J. (1988). A practitioner's guide to computation and interpretation of reliability indices for mastery tests. *Journal of Educational Measurement, 25*, 47-55.

Umesh, U. N., Peterson, R. A., & Sauber, M. H. (1989). Interjudge agreement and the maximum value of kappa. *Educational and Psychological Measurement, 49*, 835-850.

Walker de Felix, J., Waxman, H., Paige, S., & Huang, S. Y. L. (1993). A comparison of classroom instruction in bilingual and monolingual secondary school classrooms. *Peabody Journal of Education, 69*, 102-116.

Zwick, R. (1988). Another look at interrater agreement. *Psychological Bulletin, 103*, 374-378.

Send correspondence to: David A.Walker
Northern Illinois University
Email: dawalker@niu.edu

POSTMASTER: Send address changes to:
Cynthia Campbell, Managing Editor
Department of Educational Technology, Research, & Assessment
Northern Illinois University
DeKalb, IL 60115-2854

Multiple Linear Regression Viewpoints (ISSN 0195-7171) is published by the AERA Special Interest Group on Multiple Linear Regression: General Linear Model through **Northern Illinois University** and the **University of Alabama-Birmingham**.