Psychological Methods
2009, Vol. 14, No. 4, 349–366

1082-989X/09/$12.00 DOI: 10.1037/a0016956

A Conceptual and Empirical Examination of Justifications for Dichotomization

Jamie DeCoster
Institute for Social Science Research, University of Alabama

Anne-Marie R. Iselin
University of Pittsburgh School of Medicine

Marcello Gallucci
The University of Milano–Bicocca

Despite many articles reporting the problems of dichotomizing continuous measures, researchers still commonly use this practice. The authors' purpose in this article was to understand the reasons that people still dichotomize and to determine whether any of these reasons are valid. They contacted 66 researchers who had published articles using dichotomized variables and obtained their justifications for dichotomization. They also contacted 53 authors of articles published in *Psychological Methods* and asked them to identify any situations in which they believed dichotomized indicators could perform better. Justifications provided by these two groups fell into three broad categories, which the authors explored both logically and with Monte Carlo simulations. Continuous indicators were superior in the majority of circumstances and never performed substantially worse than the dichotomized indicators, but the simulations did reveal specific situations in which dichotomized indicators performed as well as or better than the original continuous indictors. The authors also considered several justifications for dichotomization that did not lend themselves to simulation, but in each case they found compelling arguments to address these situations using techniques other than dichotomization.

Keywords: dichotomization, continuous measures, dimensional measures, median splits, Monte Carlo simulations

> There are two kinds of people in the world, those who believe there are two kinds of people in the world and those who don't.—Robert Benchley, *Benchley's Law of Distinction*

Dichotomization is a statistical procedure by which a variable that originally was continuous is transformed into a categorical variable based on where people fall relative to a cutoff point (Cohen, 1983). For example, a continuous measurement of test performance can be used to create "pass" and "fail" groups on the basis of whether scores are greater than or less than a relevant cutoff score, such as 65%. The practice of dichotomization has been attacked by methodologists on a number of different grounds (e.g., Cohen, 1983; Fitzsimons, 2008; Humphreys, 1978; Humphreys & Fleishman, 1974; MacCallum, Zhang, Preacher, & Rucker, 2002; Maxwell & Delaney, 1993; Maxwell, Delaney, & Dill, 1984). These methodologists have argued that it leads to a loss in analytic power and, in some cases, can create falsely significant results. These warnings, however, have not stopped researchers from dichotomizing their continuous variables prior to analysis. MacCallum et al. (2002) found that 11.5% of the articles in two top-tier journals (*Journal of Consulting and Clinical Psychology* and *Journal of Personality and Social Psychology*) contained analyses in which at least one continuous variable was artificially dichotomized. If dichotomization is known to be problematic, why is it still used in research? The disconnect between statistical recommendations and actual practice suggests that methodologists have not fully addressed the questions that researchers have about dichotomization. Our goals in the

Jamie DeCoster, Institute for Social Science Research, University of Alabama; Anne-Marie R. Iselin, Western Psychiatric Institute and Clinic, University of Pittsburgh School of Medicine; Marcello Gallucci, Department of Psychology, University of Milano–Bicocca, Milano, Italy.

We would like to thank Sander Koole for help in the initial conceptualization of this project. We would also like to thank Kenny Lichstein, Randy Salekin, and Lynn Snow for comments made on earlier versions of this article.

Correspondence concerning this article should be addressed to Jamie DeCoster, Box 870216, Institute for Social Science Research, University of Alabama, Tuscaloosa, AL 35487. E-mail: jamie@ua.edu

current article were therefore to determine why researchers still choose to dichotomize, to thoroughly examine the validity of these reasons, and to provide recommendations for future analytic practice.

We cast our discussion at two different audiences. First, we wanted to provide the general research community with specific information regarding the circumstances under which the dichotomization of continuous variables may be justified. We hoped to resolve any questions in researchers' minds regarding whether a particular reason for dichotomization is valid or not so that they can perform the most powerful and appropriate statistical tests. Second, we hoped to provide a resource for statisticians and methodologists who collaborate or consult with investigators who might wish to dichotomize. Although many members of this second audience are aware of the problems with dichotomization and avoid it in their own work, our evaluation of the common justifications for dichotomization will allow them to make stronger, empirically based arguments against the practice in circumstances in which it leads to less valid results.

A Brief History of Dichotomization

Although the practice of dichotomization is much criticized by today's statisticians, it derives from a previously accepted tradition of categorizing responses to save labor in statistical calculations (Cohen, 1983). When it was still common to perform statistical analyses by hand, some researchers would choose to aggregate groups of observations to reduce the number of values that had to be entered into their computations. It was well understood, however, that this aggregation would reduce the variability in the measures and correspondingly weaken the estimated size of the relations between variables. Peters and Van Voorhis (1940, p. 398) showed that two-group splits will reduce correlations by 20.2%, three-group splits will reduce correlations by 14.1%, and four-group splits will reduce correlations by 8.5%, assuming there are equal numbers in each group. This is actually the best that categorization can do—if researchers choose to use groups of unequal sizes, then the observed relations will be reduced even further (Cohen, 1983). As a practical example, if the original correlation between two continuous variables was .500, we would expect to observe a correlation of .399 if one of the variables was dichotomized and a correlation of .318 if both of the variables were dichotomized, assuming that the dichotomized group sizes were equal. Peters and Van Voorhis (1940) specifically suggested that researchers should always correct for this reduction when they artificially categorize data into six or fewer groups.

Methodologists started criticizing the practice of categorizing continuous data when researchers began to use it for reasons other than ease of computation, and when authors failed to consider the effects of categorization on their results. Humphreys and Fleishman (1974) and Humphreys (1978) criticized the common use of dichotomization to allow researchers to analyze the influence of continuously measured personality variables using analysis of variance (ANOVA). These authors suggested that the use of ANOVA in this circumstance misrepresents the relations among variables found in the real world, gives the illusion of experimental control to designs that lack it, and reduces the size of the observed relations. They proposed that continuously measured variables should instead be left in their original form and be investigated with correlations and regression analysis.

The issues surrounding dichotomizing naturally continuous variables were summarized and brought to the psychological literature by Cohen (1983). He discussed several of the reasons that researchers typically give for dichotomization (e.g., that it makes analyses easier to conduct and interpret, that it allows the use of statistical techniques such as ANOVA and log-linear modeling, and that it refines crude measurements). His conclusion was that these represented either misunderstandings of statistical principles (such as in the case of refining crude measurements, because dichotomization actually adds errors of discreteness to existing measurement error) or else were not worth the costs in terms of statistical power and accuracy of the estimated relations.

MacCallum et al. (2002) provided a thorough review of the problems of dichotomization and provided a practical explanation as to why it reduces the observed relations among variables. Compared with the original continuous measure, a dichotomized variable is less precise because it does not allow the researcher to discriminate between differently scoring members in the same group. For example, someone who is just barely above the cutoff value is treated the same as someone who is near the maximum value on the scale. All of the information that distinguishes an observation from other members of its group is necessarily lost—essentially, everyone within the group is treated as having a value equal to the group mean. Losing this information makes it more difficult to use the dichotomized variable to predict participants' characteristics on other measures. Tests based on dichotomized variables will therefore have less power than those performed with the original continuous measures. Effect size estimates based on dichotomized scores will also typically be smaller than those based on the original continuous measures.

In addition to reducing the ability to detect relations, Maxwell and Delaney (1993) noted that the use of dichotomized measures can lead to spuriously significant results when two artificially dichotomized independent variables (IVs) are used to predict a dependent variable (DV) in a multifactor ANOVA. Such results can occur when the constructs underlying the two dichotomized IVs are correlated,

and only one of them is actually related to the DV. We have already discussed how it would be more difficult to relate a dichotomized variable to other measures because of the loss of information. One implication of the reduced predictive ability of the dichotomized measure is that it cannot explain as much variability in the DV as the original continuous measure. This means that there will be some variability in the DV related to the theoretical construct underlying the dichotomized measure that the dichotomized measure itself cannot explain. Now consider a third variable that has no independent relation with the DV but is related to the construct underlying the dichotomized measure. If we add this third variable to our model, this variable may be given "credit" for explaining this leftover variability. More generally, Maxwell and Delaney (1993) have mathematically shown that this can lead to inflated Type I error rates whenever there is a correlation between two dichotomized IVs. They also showed that artificially dichotomizing IVs will lead to an inflation of the Type I error rates for the test of an interaction between those variables if the IVs are correlated and one of the IVs has a nonlinear relation with the DV. Vargha, Rudas, Delaney, and Maxwell (1996) further showed that the spuriously significant results discussed by Maxwell and Delaney will also occur when only one of the two IVs is dichotomized. Taken together, these findings indicate that dichotomizing variables can not only reduce the power of statistical tests but can also lead to incorrectly significant results.

Why Do Researchers Dichotomize Their Data?

Most researchers are aware of the problems that may result from dichotomization, and yet dichotomized variables can still be regularly found in current research (Fitzsimons, 2008). In their discussion of the prevalence of dichotomization, MacCallum et al. (2002) identified 105 articles published in either the *Journal of Consulting and Clinical Psychology* or the *Journal of Personality and Social Psychology* over a 3-year period that included analyses with at least one artificially dichotomized continuous variable. To obtain a better understanding of why researchers use dichotomization, we e-mailed the contact authors for these articles and asked them to explain why they had dichotomized variables in their studies and to identify any circumstances in which they thought dichotomization might be appropriate. Most of the research examining the performance of dichotomized variables has focused on idealized, tractable variables. We thought that researchers with practical experience in dichotomizing variables might be aware of specific circumstances that have not been investigated in formal analyses in which dichotomization is preferable. Sixty-six (63%) of these authors responded to our query, providing us with a survey of modern researchers' beliefs about dichotomization.

We also sought the opinions of researchers with strong methodological and statistical backgrounds, believing that these individuals might be aware of additional situations in which dichotomized variables might be optimal. To accomplish this, we e-mailed the contact authors of all articles published in *Psychological Methods* between 2003 and 2008. We asked these individuals whether they had used dichotomization in their own research and whether they could identify situations in which they thought a dichotomized indicator could outperform a continuous indicator (i.e., be more powerful or accurate). Fifty-three (40%) of these authors responded to our query.

Table 1 presents a summary of the possible reasons to use dichotomization that were provided by the authors responding to our e-mails. We were able to categorize these reasons into three major groups. The first suggested that there were particular types of variables or particular types of relations between variables that might be better examined with dichotomized indicators. The second suggested that conducting analyses with dichotomized indicators is easier than conducting analysis with continuous indicators. The third suggested that under certain circumstances, the analyses conducted with dichotomized indicators may better match the theoretical purpose of the research.

In the following sections, we examine each of these reasons individually and explain why the authors felt that it might offer a justification for dichotomization. We then provide a critical examination of the reason to determine whether it is consistent with current statistical knowledge. We include Monte Carlo simulations to empirically investigate reasons related to the distributions of the variables. The simulations share several common properties, so we present them together at the end of the section on distributions to make understanding their procedures and results simpler.

In our discussions, we will use the term *latent variable* to refer to the theoretical true score on a construct, which in practice cannot be measured. The term *observed variable* refers to the score that is obtained on a continuous measure designed to capture the latent variable. An *indicator variable* is an index derived from the observed variable that is actually used in analysis. We most commonly will be considering either *continuous indicators*, which would be exactly equal to the observed variable, or *dichotomized indicators*, which would be categorical variables created by dichotomizing the observed variable (such as by performing a median split). As an example, a researcher may be interested in the latent variable of intelligence, which is in practice estimated using the observed measure of an IQ test. When analyzing relations with intelligence, the researcher might use the IQ score in its raw form as a continuous indicator or choose to perform a median split to create a dichotomized indicator representing "high IQ" and "low IQ" groups.

Table 1
Reasons for Dichotomization Offered by Authors

Reasons cited for dichotomization	Practitioners of dichotomization mentioning reason ($N = 66$)	*Psychological Methods* authors mentioning reason ($N = 53$)
Reasons related to the distributions of the variables		
The latent variable has an irregular distribution (with the irregularity unspecified).	0	4
The latent variable being measured is truly categorical.	10	14
The latent variable is skewed.	1	5
The observed variable has poor reliability.	8	6
The observed variable has outliers.	2	1
The study uses extreme group analysis.	5	4
The relation between the latent and outcome variables is not linear.	7	13
Reasons related to the ease of analysis		
Results from analyses with dichotomized variables typically lead to the same conclusions as those with continuous variables.	5	1
It is easier to present the results from analyses with dichotomized IVs.	10	4
It is easier to analyze interactions with dichotomized IVs.	6	2
Reasons related to the prior use of the variable		
The field has identified theoretically meaningful cut points on the variable being dichotomized.	13	16
Researchers have typically dichotomized the variable in the past.	0	4
Authors stating that dichotomization is never justified	1	10

Note. IVs = independent variables.

Reasons Related to the Distributions of the Variables

The latent variable has an irregular distribution. While methodologists have shown through mathematical proofs and Monte Carlo simulations that dichotomization leads to weaker and potentially misleading statistical tests, almost all of these were based on the assumption that the variable being dichotomized has a naturally continuous and typically normal distribution (e.g., Cohen, 1983; MacCallum et al., 2002; Maxwell & Delaney, 1993). This leaves the unanswered question of what happens when researchers dichotomize variables that have irregular distributions. It is possible that there are specific types of distributions in which dichotomized indicators provide better representations of the underlying constructs than the observed continuous indicators.

Researchers most commonly reported that dichotomization might be appropriate when the underlying construct being measured is truly categorical. MacCallum et al. (2002) specifically stated that one of the circumstances under which it might be appropriate to use dichotomization is when the researcher believes that the underlying construct has a categorical structure. The rationale is that a dichotomization of the observed measure more naturally reflects the latent construct than the observed continuous measure. Without a doubt, the distribution of the dichotomized indicator appears to be more like that of the latent variable than the continuous indicator. However, this does not guarantee that the observations are assigned to the correct groups and makes the cost of mismatches more extreme since there is no middle ground. A number of researchers also mentioned the possibility that a dichotomized indicator might perform better when the underlying latent variable is highly skewed. It is possible to conceptualize a skewed distribution as consisting of two parts: the observations surrounding the distribution's mode and the observations in the distribution's tail. Sometimes researchers believe that these parts represent theoretically different cases and feel that a dichotomization of the variable would represent the variable better than the original continuous score.

The observed variable has poor reliability. The reliability of a variable is a quantitative estimate of how much random error is incorporated in its measurement, such that measures with lower reliabilities have a greater proportion of random error (DeVellis, 2003). The authors that we surveyed believed that low reliability could provide a justification for dichotomization because continuous variables make finer distinctions between individuals than categorical variables and so should be more affected by random error in the measurements. They suggested that dichotomization

decreases variability in the responses (because scores are collapsed within each group), reducing the random error and making the results more accurate.

The observed variable has outliers. By definition, continuous measures can accommodate a wider range of responses than dichotomized measures, which also means that the potential impact of unusual or erroneous responses is greater for continuous measures. It has been well established that such outliers can drastically change the characteristics of an estimated regression line (Neter, Kutner, Nachtsheim, & Wasserman, 1996). Even when they are consistent with the regression line fitted through the other data points, observations that have extreme values on the IV dramatically affect tests of the regression models. After dichotomization, however, all values on the same side of the cutoff point have exactly the same influence on the statistical results. Some researchers therefore argue that dichotomization should be used to prevent outliers from strongly biasing statistical tests.

The study uses extreme group analysis. Sometimes researchers will choose to analyze only those people who have extreme values on a variable, either by not collecting data from or by excluding the results from those who have moderate scores on the variable. It has been known for some time that it is possible to increase the power of tests relating a continuous IV to a DV by recruiting people from the extreme ends of the distribution on the IV (Feldt, 1961). By excluding people from the middle of the distribution, a procedure called *extreme groups analysis*, researchers increase the variability within the sample, which in turn leads to stronger observed relations between the IV and the DV. This method has been shown to be an effective way of selecting participants to maximize the likelihood of finding a significant relation between an IV and a DV (Preacher, Rucker, MacCallum, & Nicewander, 2005).

Once the high-scoring and low-scoring individuals have been selected from the distribution, it is common practice to dichotomize the values and create a categorical variable with two groups. Extreme group analysis forces the distribution of the variable to become more categorical, which can be used as a justification for dichotomization. While there is still variability within the upper and lower parts of the distribution following this procedure, this is typically much smaller than the variability between the groups, possibly making dichotomization more appropriate than a variable containing the full range of values.

The relation between the latent and outcome variables is not linear. The primary tool used to analyze the influence of a continuous IV on a DV is linear regression, which assumes that the effect of changes in the IV is constant across the entire length of the scale. Sometimes, however, researchers may expect that there are certain thresholds in which even a small change in the value of the IV can have a substantial influence on the DV, even while changes in the IV have minimal influences on the DV at other points of the scale. For example, a particular drug may have little effect until the dosage reaches a certain threshold, at which point it has a substantial effect that is constant for any dosage greater than the threshold. This type of a relation may possibly be best tested with a dichotomization of the IV that separates participants who fall below the threshold from those who fall above it.

Simulations Investigating the Effects of Distributional Characteristics on Dichotomization

We performed simulations to examine how continuous and dichotomized indicators perform as we changed characteristics of the latent, observed, and outcome variables. We conducted three primary sets of simulations. In the first set, we investigated whether the continuous observed variable or a dichotomization of this variable had stronger relations with the underlying latent variable. In these simulations, we varied the latent variable continuity, latent variable skewness, and observed variable reliability as well as whether we performed the analysis using the full data set or only extreme groups. In the second set of simulations, we compared the performance of continuous and dichotomized indicators when there were outliers in the distribution of the observed variable. In the third set of simulations, we investigated the effects of changing the nature of the relation between the latent and outcome variables (i.e., whether it was linear or nonlinear).

In our simulations, we assumed that the measurement of psychological characteristics could be assigned numerically meaningful values. Michell (1997) discussed the implications of this assumption, noting that there are alternative ways to interpret these measured characteristics. The validity of the conclusions drawn from our simulations therefore depends on the validity of conceptualizing psychological measurements as real numbers. Past researchers studying dichotomization (e.g., MacCallum et al., 2002) have also made this assumption, and there is a long tradition of treating measured psychological characteristics in this manner, going back to Fechner (1860). We therefore believe that making this assumption is reasonably justified and that valid information can be obtained by examining the results of our simulations.

Generating latent, observed, and indicator variables. We determined the latent, observed, and indicator variables the same way for all of our simulations. We first generated the distribution of the underlying latent variable. We wanted to do this in a way that would let us vary both the latent variable's continuity (allowing it to be either truly continuous or truly dichotomous) and skewness (allowing it to be skewed or unskewed in the continuous case, or have groups of equal or unequal size in the categorical case). Using a

specialized version of the generalized logistic function (Richards, 1959), we were able to randomly generate latent distributions that independently varied on these two dimensions. The value of the latent variable η_i for a given observation was determined using the following formula:

$$\eta_i = \frac{1}{1 + e^{-B(x_i - M)}},$$

where x_i is a random variable uniformly distributed between 0 and 1, B is the continuity parameter, and M is the skewness parameter. B can take values between 1 and positive infinity, such that higher values correspond to more categorical distributions. M can take values between 0 and 1, such that a value of .5 corresponds to a symmetric distribution, values close to 0 correspond to left-skewed distributions, and values close to 1 correspond to right-skewed distributions. In our simulations, we only considered right-skewed distributions because there is no reason to expect that the direction of skewness would influence its effects.

The way we used this formula to generate our latent distribution was to first set B and M to the values specified by the simulation we wanted to conduct. We then randomly generated 500 numbers between 0 and 1 to be the values of x_i. Finally, we computed the values of η_i by substituting the values of x_i into the generalized logistic function. This gave us the 500 values for our latent distribution, representing the true values on the latent construct for all of the participants in a single simulated study. Details of the mathematical model used to generate data for our simulations are summarized in the Appendix.

Examples of different latent distributions obtained using different values of B and M are presented in Figure 1.

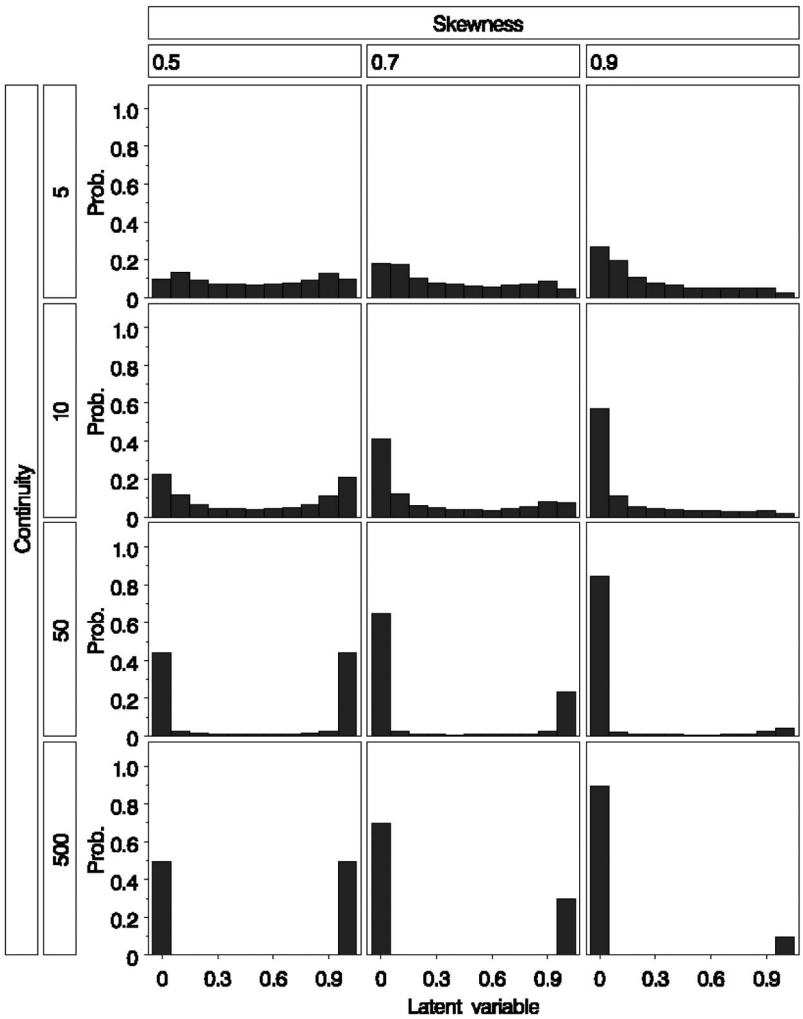

Figure 1. Latent distributions obtained using different values of continuity and skewness. Prob. = probability.

The effects of the continuity parameter B can be seen by comparing the distributions across the different rows. At the lowest level of B, the distribution gradually goes from 0 to 1, providing us with a continuous distribution. As B increases, the observations become more and more clustered at the endpoints, making the distribution more categorical. The effects of the skewness parameter M can be seen by comparing the distributions across the columns. When this parameter is equal to .50, we have an approximately flat distribution in the continuous case and equally sized groups in the categorical case. As this parameter gets closer to 1, we see that the continuous distribution becomes more skewed and the categorical distribution becomes more unbalanced. If we had chosen to examine values for M that moved toward 0 instead of those that moved toward 1, we would have obtained an exact mirror image of the results displayed in this figure.

After the values of the latent variables were determined, we computed the values of the continuous observed variable m_i, for each case using the equation

$$m_i = \left(\frac{R}{\sqrt{1 - R^2}}\right)\eta_i^* + e_i,$$

where η_i^* is the standardized value of the latent variable for case i, e_i is a random value chosen from the standardized normal distribution, and R is the reliability of the outcome measure. The expected correlation between the observed variable m_i and the latent variable η_i is exactly R (cf. Appendix).

The final distribution of the observed variable was therefore determined by three parameters: the continuity parameter of the latent variable (B), the skewness parameter of the latent variable (M), and the reliability of the observed variable (R). An important benefit of using this method to determine our distributions is that we could independently vary each of these dimensions in a continuous fashion and observe their incremental influences on continuous and dichotomized indicators. Examining gradations of continuity, skewness, and reliability in this way allows a fuller understanding of how changes in these features influence the performance of continuous and dichotomized indicators than we would have obtained from simulations that examined them in an all-or-none fashion.

Most of our simulations derived three different indicator variables from the original observed variable.

1. A *continuous indicator*, which is simply equal to the value of the observed variable.

2. A *median split*, which is a categorical variable in which individuals in the bottom half of the distribution of the observed variable are assigned the value of 0, and the remainder are assigned the value of 1.

3. A *proportional split*, which is a categorical variable in which the split is determined by the value of the skewness parameter. Specifically, individuals who score in the lower $M * 100\%$ of the distribution of the observed variable are assigned the value of 0, and the remainder are assigned the value of 1. For example, if the value of M is .7, then a proportional split would put 70% of the observations in the lower group and 30% of the observations in the upper group. With a proportional split, more observations are placed in the lower group when more observations are found in the lower part of the latent distribution. If the latent variable is categorical, the proportions assigned to each group of the indicator duplicate the proportions found in the latent distribution.

Examining these three indicators allowed us both to compare the performances of continuous and dichotomized indicators as well as to see how choosing a nonproportional split affects the performance of dichotomized indicators.

Simulation 1: Continuity, skewness, reliability, and extreme group analysis. Our first set of simulations explored how dichotomization influenced the ability of continuous and dichotomized indicators to accurately represent the latent variable under different conditions of continuity, skewness, and reliability. For these simulations, we determined the fit of each indicator by measuring its correlation with the latent variable. We conducted a total of 2,400 simulations: 50 in each combination of four levels of continuity (B = 5, 10, 50, 500), three levels of skewness (M = .50, .70, .90), and four levels of reliability (R = .40, .60, .80, .90). For each simulation, we randomly determined values of the latent variable for 500 subjects using the generalized logistic function, combined the latent variable with random error to obtain values of the observed distribution, and then used the observed distribution to compute continuous, median split, and proportional split indicators. We then correlated the values of the latent variable with the values of the various indicators. Each simulation was therefore conceptually parallel to collecting data from a single study. We conducted multiple simulations within each condition to ensure that our results represented overall trends. We did not base our conclusions on statistical tests because with 2,400 studies, each including data from 500 participants, every comparison was statistically significant. We instead chose to present the data graphically and base our conclusions on the broader trends that were apparent in the figures. We used Fisher's r-to-Z transformation when averaging the

correlations across multiple simulations and then converted these averages back to correlations using Fisher's *Z*-to-*r* transformation.[1]

Figure 2 presents the average correlations between the different indicators and the underlying latent variable separately for each level of continuity. Higher correlations denote that an indicator provided a better representation of the underlying latent variable. This graph shows that the continuous indicator performed equally well whether the latent distribution was categorical or continuous and performed substantially better than either dichotomized indicator when the latent distribution was more continuous. As the latent distribution became more categorical, the performance of the proportional split indicator improved until it was equal to that of the continuous indicator. The median split, however, never performed as well as the continuous indicator.

These results indicate that when the latent variable was truly categorical, both continuous and proportional split indicators performed equally well. However, as the distribution of the latent variable became more continuous, the performance of the dichotomized indicators worsened while the performance of the continuous indicator stayed the same. These results suggest that the continuous indicator is more robust than the dichotomized indicators, in that it can perform well with any type of distribution. The results also suggest, however, that a proportional split can perform just as well when the latent distribution is truly categorical.

Figure 3 displays the average correlation between each type of indicator with the latent variable separately for different combinations of continuity and skewness. The graph on the left shows the performance of the continuous and median split indicators when the latent distribution was symmetric. This graph only has two lines because the proportional split and the median split created exactly the same groups when there is no skewness. In this case, we see an advantage of the continuous indicator over the dichotomized indicator, although this advantage decreases as the latent distribution becomes more categorical. The graph on the right illustrates the results when the latent distribution was skewed. In this case, the continuous indicator performed better than either of the dichotomized indicators when the latent variable was continuous. As the latent distribution became more categorical, the performance of the proportional split consistently improved, exceeding the performance of the continuous indicator when the distribution was entirely categorical with skewed data.

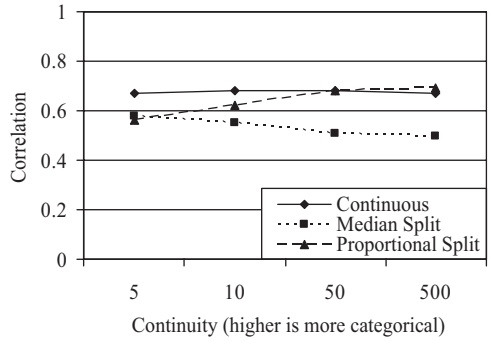

Figure 2. Average correlations with the latent variable by continuity and indicator type. The vertical axis represents the mean correlation of the indicator with the original latent variable.

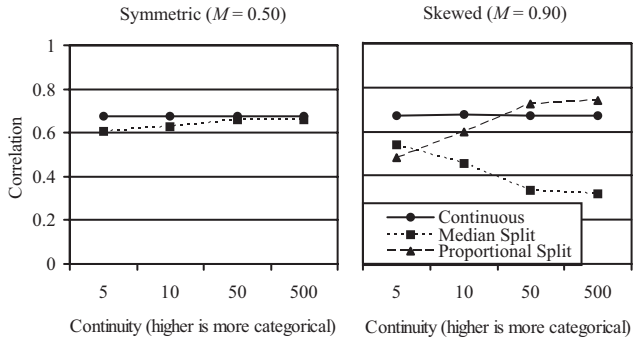

Figure 3. Average correlations with the latent variable by continuity, skewness, and indicator type. The vertical axis represents the mean correlation of the indicator with the original latent variable.

These results indicate that although dichotomized indicators are viable alternatives to continuous indicators when the latent distribution is categorical, this is only true when the relative sizes of the dichotomized groups match those found in the underlying distribution. When the latent variable is skewed, the proportional split indicator shows improved performance as the latent distribution becomes more categorical, but the median split does not. When the underlying distribution is fully or moderately continuous, a dichotomized indicator cannot perform at the same level as the continuous indicator, even with appropriately matched skewness.

Figure 4 displays the average correlation of each type of indicator with the latent variable separately for different combinations of continuity and reliability. When the latent distribution was truly linear, the continuous indicator outperformed the dichotomized indicator across all levels of reliability. There was also an influence of reliability on the performance of the indicators when the latent distribution was categorical, such that the continuous indicator was

[1] Schulze (2004) noted that using Fisher's *r*-to-*Z* transformation can increase the bias of correlation coefficients. However, we felt that using this transformation was preferable when averaging correlations over simulations because the bounded nature of the correlation coefficient makes its distribution asymmetric, and the bias in *Zr* is often considered to be small enough to be safely ignored (Snedecor & Cochron, 1989).

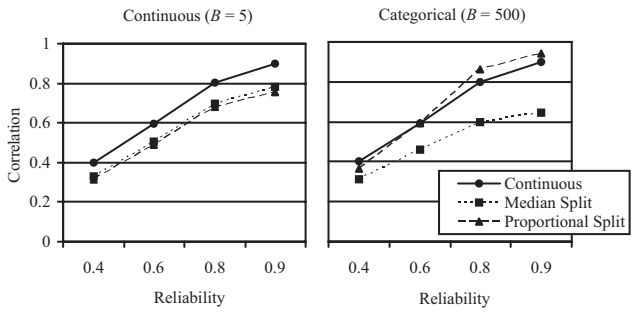

Figure 4. Average correlations with the latent variable by reliability, continuity, and indicator type. The vertical axis represents the mean correlation of the indicator with the original latent variable.

superior at low levels of reliability, but the proportional split indicator was superior at high levels of reliability.

These results differ from the expectations of researchers responding to our surveys in two ways. First, dichotomized indicators do not perform better under conditions of poor reliability. This is because dichotomization is necessarily applied after the original continuous measurement has been obtained, which means that the random error has already influenced the results. Although dichotomization will reduce the amount of random error in the data, this reduction is directly proportional to the reduction of systematic variability in the measurement. It is the relative proportion of systematic to random variability that is important for test statistics, so dichotomization does not provide a more precise measure of the underlying construct. It also comes with a reduction in the total variability of the measure, which inhibits the ability of the measure to relate to other variables (Cohen, Cohen, West, & Aiken, 2003).

Second, when the latent variable is naturally categorical, the proportional split indicator outperforms the continuous indicator when the reliability is high. We believe that dichotomized indicators can outperform continuous indicators when the latent distribution is categorical and the amount of random error in the measurement is small relative to the difference between the two group means. In this case, deviations from the group mean provide no information about the true values of the individual, and these deviations are small enough that very few errors are made when dichotomization is applied. It is also interesting to note that the advantage of the proportional split indicator declines when moving from a reliability of .8 to a reliability of .9. In truly categorical distributions, the values of the continuous and proportional split indicators become more similar as the reliability increases, until they are exactly the same when reliability = 1.0. At this point, both indicators exactly replicate the underlying latent distribution. The difference between the continuous and proportional split indicators will decrease at the highest levels of reliability, which will correspondingly decrease the advantage of the proportional split indicator.

It is important to distinguish dichotomizing a measure after it has been collected from limiting the number of response options in a scale as it is administered. A number of studies have shown that under certain circumstances, scales with a smaller number of response options can produce more valid and reliable data than those with a larger number of response options (e.g., Matell & Jacoby, 1971; Miethe, 1985; Wikman & Warneryd, 1990). The fact that a dichotomous response may at times be more appropriate than a continuous response is not under debate. The question being considered is whether dichotomization can help refine a noisy measurement after it has already been collected on a continuous scale. It is this latter possibility that fails to receive support from our simulations.

We based our analyses of the effect of extreme group analysis on these same simulations, comparing the results found using all of the participants to those found when we excluded participants who scored between the 25th and 75th percentile of the latent distribution. However, we only considered simulations that had symmetric distributions (i.e., skewness parameter $M = .50$). Dropping observations from a nonsymmetrical distribution affects many things other than simply the extremity of the observations, which would complicate the interpretation of the results. Excluding observations between the 25th and 75th percentile of a skewed distribution creates dichotomized groups that differ in terms of their (a) variance, (b) average difference from the overall grand mean, and (c) distribution shape. Any differences observed between the indicators could therefore be due to these factors instead of the effect of extreme group analysis. In symmetric distributions, the proportional split would be a median split, so we only needed to consider continuous and median split indicators. Restricting our attention to symmetric distributions does limit the generalizability of our conclusions to these situations. However, the central limit theorem suggests that emergent characteristics that arise from linear combinations of other variables tend to follow a normal (and therefore symmetric) distribution (Hays, 1994, pp. 243–244). Given the prevalence of such characteristics in nature, we feel that our results still have broad applicability.

Figure 5 displays the average correlations of the continuous and median split indicators at each level of continuity with and without extreme group analysis. The relations for both types of indicators appeared to be stronger for extreme group analysis than for analyses in which the full range of data was used. Comparing the indicators, we can see that the median split performed worse than the continuous indicator when the full range of data was used but that the two indicators represented the latent variable equally well when extreme group analysis was used. Although Alf and Abrahams (1975) have shown empirically that analyses that treat

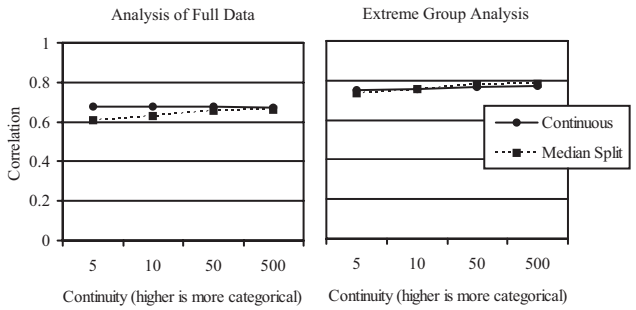

Figure 5. Average correlations with the latent variable by continuity, use of extreme group analysis, and indicator type. The vertical axis represents the mean correlation of the indicator with the original latent variable.

the extremitized variable continuously are more powerful than those that treat it dichotomously, this advantage did not appear to be large enough to produce an observable effect in our simulations. It appears that the reduction in variability caused by extreme group analysis removed the benefit of a continuous indicator, even when the latent distribution was naturally continuous. These results suggest that continuous and dichotomized indicators may be equally viable in the performance of extreme group analysis.

Simulation 2: Outliers. The most appropriate way to handle an outlier depends on why the observation is unusual (Barnett & Lewis, 1994). Sometimes outlying observations occur because of simple typographical errors. In these cases, the appropriate action is to change the value of the outlier. At times, an outlier could represent an observation that is not a member of the population being studied (i.e., a 20-year-old accidentally recruited in a study of older adults). In these cases, the appropriate action is to remove the observation from the analysis. At other times, outliers may simply represent unusual individuals or events that still fit within the population of interest. In these cases, the researchers may keep the observation after transforming the data, truncating the responses, or using procedures robust to the presence of outliers. Rather than choosing to dichotomize, in which all outliers are treated in exactly the same way, researchers are better off intentionally looking for outliers and then handling each in an individualized fashion.

All of this being said, we decided to conduct simulations to determine the extent to which dichotomization reduces the impact of outliers in a distribution. We compared dichotomization to both the original continuous distribution as well as the continuous distribution after Winsorizing the data (Ruppert, 1988), in which observations below the 5th percentile of the distribution are set equal to the value at the 5th percentile and observations above the 95th percentile of the distribution are set equal to the value at the 95th percentile.

We used Monte Carlo simulations to explore the effects of outliers on continuous, dichotomized, and Winsorized indicators. In each simulation, 95% of the observations were considered to be typical observations, and 5% were considered to be outliers. The outliers were drawn from a distribution that was the same as the distribution of the typical observations in all characteristics except for the mean, which was at least 2 standard deviations greater than the mean of the typical observations. We conducted a total of 600 simulations, 50 in each combination of four levels of continuity (B = 5, 10, 50, 500) with three levels of outlier extremity (outlier M = 2, 4, or 6 standard deviations greater). We limited our consideration to symmetric distributions to make the effect of outliers more obvious. This also removed the need to separately consider proportional split indicators because the proportional split is a median split in symmetric distributions.

Figure 6 displays the average correlations of the three different indicators with the outcome variable separately by continuity and outlier extremity. When the outliers were only slightly deviant from the original distribution (with a mean 2 standard deviations higher), the continuous indicator performed equal to or better than the dichotomized indicator across the range of continuity, while the Winsorized indicator performed better than both. When the outliers were more extreme (with a mean 6 standard deviations higher than the original distribution), there is a clear advantage of the dichotomized indicator over the continuous indicator at all levels of continuity. The dichotomized indicator, however, performed notably worse than the Winsorized indicator. These results suggest that the fact that dichotomization controls the effect of extreme outliers is not a sufficient justification for dichotomization because Winsorizing the continuous indicator provides a consistently better solution.

Simulation 3: Linearity. The simulations investigating the effect of linearity were more complex because they required the additional consideration of an outcome variable and its relation to the latent variable. For each simulation, we determined the distributions of the latent, observed, and

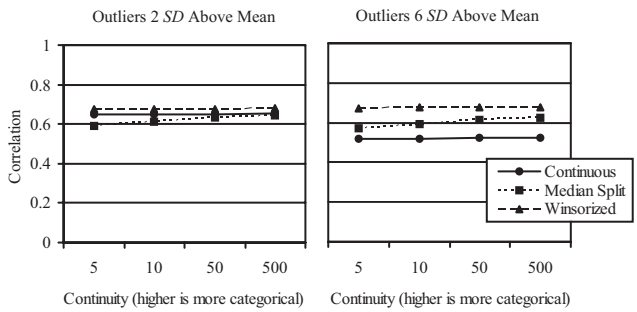

Figure 6. Average correlations with the latent variable by continuity, outlier extremity, and indicator type. The vertical axis represents the mean correlation of the indicator with the original latent variable.

indicator variables using the procedures described for Simulation 1. We then determined the value of an outcome variable, which was a function of the latent variable and random error. The nature of the relation between the latent variable and the outcome variable was determined by the equation

$$\varphi_i = \frac{1}{1 + e^{-L\left(\frac{\eta_i}{s_\eta} - .5\right)}},$$

where L is the linearity parameter, η_i is the latent variable, and s_η is the standard deviation of the latent variable. L can take values from 1 to positive infinity, and it alters the nature of the relation between the latent and outcome variables such that higher values correspond to greater deviations from linearity. An illustration of the relations obtained under different levels of L is presented in Figure 7. When $L = 1$, the relation between the variables will be linear. At moderate values of L (5 and 50 in our simulations), the relation is sigmoidal (S-shaped). At high values of L (500 in our simulations), the relation is a step function. Normally distributed random error is added to the function to make the relation probabilistic instead of deterministic, with variance appropriately chosen to fix the correlation between the latent variable and the outcome variable to a constant value (see the Appendix).

We conducted a total of 800 simulations: 50 in each combination of four levels of continuity ($B = 5, 10, 50, 500$) with four levels of linearity ($L = 1, 5, 50, 500$). For simplification of the interpretation of the results, all of the latent distributions were symmetric ($M = .50$), the observed variable always had a high but not perfect reliability ($R = .70$), and the relation between the latent and outcome variables was always strong ($r_{y\eta} = .70$). For each simulation, we randomly determined values of the latent variable for 500 subjects based on the appropriate level of continuity, combined the latent variable with random error to obtain values of the observed distribution, used the observed distribution to compute continuous and median split indicators, and determined the value of the outcome variable with a function based on the appropriate level of linearity. We then correlated each of the indicator variables with the value of the outcome variable. We used Fisher's r-to-Z transformation when averaging the correlations across multiple simulations and then converted these averages back to correlations using Fisher's Z-to-r transformation.

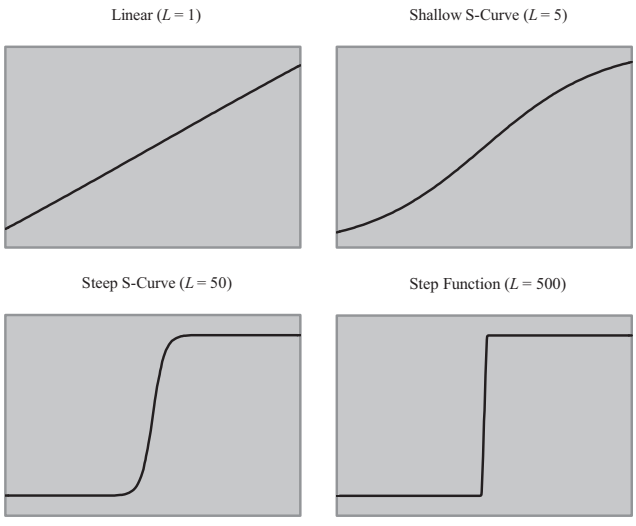

Figure 7. Relations between latent variable and outcome at different levels of linearity.

Although the linear model is a less-than-optimal way to analyze the continuous variable, we decided to test the predictive ability of the continuous indicator using such a model in our simulations because it more closely paralleled the way we analyzed a dichotomized indicator. A nonlinear equation would more accurately represent the nature of the relation between a continuous indicator and an outcome, optimizing the continuous indicator's performance. Our logic here was that if the dichotomized indicator cannot outperform the continuous indicator with the linear model, then it will certainly not be able to outperform the continuous indicator with a model that more accurately reflects the nonlinear relation.

Figure 8 displays the average correlations of continuous and median split indicators with the outcome variable separately by continuity and relation form. In each of the graphs, the continuous indicator outperformed the median split when the latent distribution was continuous, but this difference disappeared as the distribution became more categorical. We did not see a substantial effect of linearity: The performances of both indicators appeared to be consistent across all levels of linearity, with the advantage of the continuous indicator being consistently larger when the latent variable was continuous.

Dichotomization does not produce an advantage when trying to detect nonlinear relations because any nonlinear relation that can be detected as a difference between dichotomized group means can also be detected by linear regression. A straight line can always be fit between any two points, so linear regression can represent any relation that is identified as a difference between the mean levels of a dichotomized variable. When the mean of the lower group of a dichotomized IV is less than the mean of the upper group, the slope between the continuous IV and the DV will be greater than 0. When the mean of the lower group of a dichotomized IV is greater than the mean of the upper group, the slope between the continuous IV and the DV will be less than 0. A categorical transformation of a continuous variable must create three or more groups to represent a relation that cannot be detected with simple linear regression. Even in these cases, the proper polynomial trans-

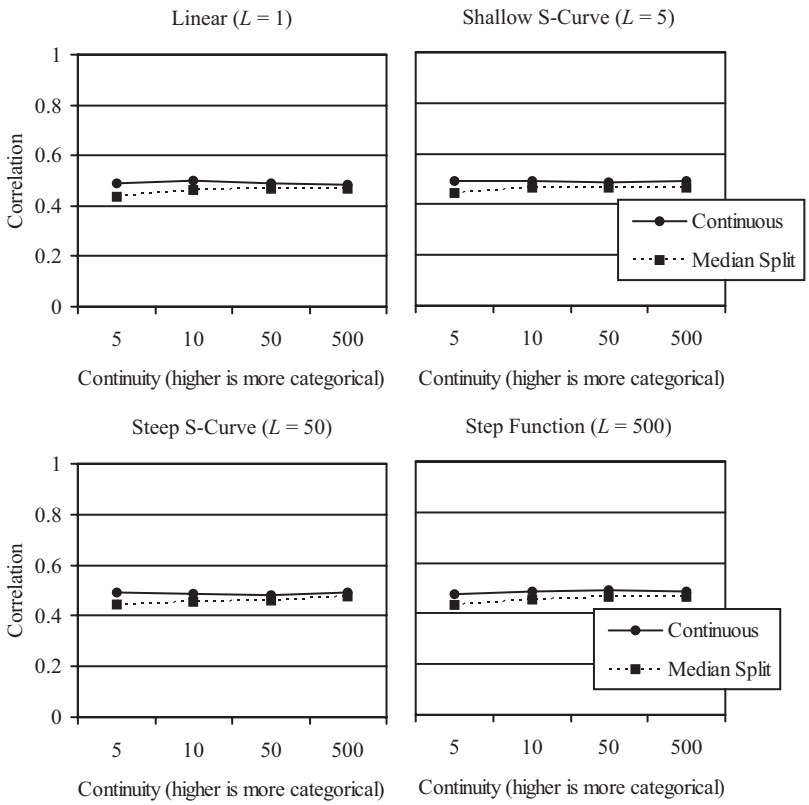

Figure 8. Average correlations with the outcome variable by continuity, linearity, and indicator type. The vertical axis represents the mean correlation of the indicator with the outcome variable.

formation of the continuous variable could detect the same relation with higher statistical power (Cohen et al., 2003).

All of this being said, neither linear regression nor a dichotomized analysis is likely the best way to test the relation between the IV and the DV if researchers suspect that the relation is nonlinear. Statisticians have developed specific analytic procedures to accurately model nonlinear relations (Seber & Wild, 2003), which use the full information available in the continuous measure to predict the DV. These models will fit the actual relation with much greater accuracy, resulting in stronger statistical tests. In addition, most tests of nonlinear relations are able to accurately model linear relations, so there is little harm in using these procedures if the original assumption about a nonlinear relation turns out to be incorrect.

Our consideration of nonlinearity focused on the assumption that the slope of the relation between the IV and the DV varied across the levels of the IV. At high levels of *L*, our implementation of nonlinearity produced a step function, which we believe to be the type of nonlinearity that is most commonly thought of when researchers decide to dichotomize their data. However, we acknowledge that this does not capture the full range of possible nonlinear functions. We therefore do not claim that our results provide an exhaustive consideration of this matter. Our simulations do show that there is no reason to use dichotomized indicators when researchers believe that a sigmoidal or step function describes the relation of the IV to the DV.

Reasons Related to the Ease of Analysis

Results from analyses with dichotomized variables typically lead to the same conclusions as those with continuous variables. Even though this article focuses on the differences between the results found with continuous variables and those found with dichotomized variables, it is important to remember that both methods are aimed at testing the same hypotheses. The analysis of continuous variables typically uses regression, and the analysis of dichotomized variables typically uses ANOVA but in both cases, the goal is to determine whether there is a relation between the IV and the DV. In practice, the conclusions drawn from use of the two methods will usually be identical. Some researchers therefore have suggested that it ultimately does not matter which of the two methods is used. These researchers further suggested that since it does not matter which of the two methods is used, variables should be dichotomized and analyzed using ANOVA if this procedure is more familiar to the researcher than regression.

While it is true that the results from continuous and dichotomized variables typically converge, it has been shown that except in the situations we identified earlier in this article, dichotomization leads to systematically smaller relations between the IV and the DV. Although the primary emphasis in the reporting of studies is still on hypothesis testing, there is a clear movement toward more reliance on estimates of effect size (American Psychological Association, 2001). Effect sizes can be reduced through the practice of dichotomization even when the conclusions being drawn from the hypothesis tests are the same. Modern researchers often not only consider the statistical significance of an effect but also judge the importance of a phenomenon on the basis of the magnitude of its effect size. Effect sizes are fundamentally important to power analyses (Cohen, 1988), and researchers who base their sample size calculations on a study in which the variables were dichotomized may end up collecting more data than would actually be required to test their hypotheses. Furthermore, dichotomization can bias meta-analytic reviews of the literature by underestimating the magnitude of effect sizes and by introducing artifactual variation (based on whether variables were dichotomized) which could be inaccurately attributed to moderator variables (Hunter & Schmidt, 1990). Given that analyses in which continuous measures are used typically provide more accurate estimates of the relations between underlying constructs than those using dichotomized measures, researchers should present the former unless they have specific evidence that they are in one of the situations in which dichotomized indicators perform equivalently to continuous indicators.

Using dichotomized IVs makes analysis of interactions easier. Although testing main effects is just as easy whether one uses regression or ANOVA, it is somewhat more difficult to test and explain interaction effects in regression. Part of this difficulty stems from the fact that many researchers are not familiar with the methods for testing interactions in regression. Whereas the examination of interaction effects was fundamentally a part of ANOVA from its beginnings (Fisher, 1925), methods for testing interaction effects in regression did not develop until much later (Cohen, 1968). Even after these procedures were developed, a long time passed before a comprehensive treatment of testing interaction effects in regression was available to researchers (Aiken & West, 1991). As a result, most researchers have been formally trained in how to test for interactions among categorical IVs, but fewer have been trained in how to test for interactions among continuous IVs. A second source of difficulty in testing interactions involving continuous IVs is a general lack of support from statistical software. Whereas most programs (such as SPSS) automatically test for interactions among categorical IVs, interactions involving continuous IVs commonly must be set up by hand through creation of multiplicative interaction terms. Not only does this make these models less accessible to researchers, it also increases the opportunity for human error to influence the results.

Although the procedures to test interactions involving continuous IVs are less familiar to researchers, they have been fully developed. Both Aiken and West (1991) and Cohen et al. (2003) provide thorough descriptions of how to test interactions among continuous IVs as well as interactions between continuous and categorical IVs. These methods require only a program that can perform multiple regression analysis and can therefore be used with any existing statistical software package. The only complication is that the IVs may need to be transformed or recoded, and new terms representing the interactions may need to be created by multiplying values together. The work required to analyze interactions when the IVs are treated continuously is therefore not much greater than that required to analyze them categorically. The primary effort will be for those unfamiliar with these methods to learn them. This one-time investment will provide ongoing dividends in more powerful and accurate interaction tests (Maxwell & Delaney, 1993).

It is easier to present the results from analyses using dichotomized IVs. After a researcher obtains a significant result, the next thing he or she must do is explain what that result means to the audience. When the test involves the relation between a categorical IV and a continuous DV, the researcher can explain the result by simply providing the means of each of the groups along with post hoc analyses indicating which groups are significantly different from each other. The presentation of a relation between a continuous IV and a continuous DV is slightly more complicated. There are no group means to present; instead, the best that a researcher can do is present a regression line illustrating the relation. While these are reasonably well understood by the academic community, they are sometimes seen as more difficult to interpret because there are no post hoc tests specifying which values of the IV are significantly different from each other.

The difficulty in presenting the results of analyses with continuous IVs is exacerbated when the statistical model includes interaction terms. With categorical IVs, the individual cell means can be plotted, providing a visual display of how the effect of one variable changes across different levels of other variables. The strength of these changes can be explored with "simple effects tests," which are used to statistically test the differences between the levels of one factor at a specific combination of the levels of the other factors involved in the interaction. The researcher can understand the interaction by seeing where the simple effect tests are significant and where they are not (Keppel, 1991). In a similar way, interactions involving continuous IVs and DVs are typically examined by comparing the "simple slopes," which are regression equations relating one of the IVs to the DV at a particular combination of the levels of the

other IVs involved in the interaction (Aiken & West, 1991). The researcher can understand the interaction by comparing the simple slopes for one variable across different levels of the other variables involved in the interaction.

Even though parallel methods for interpreting interactions exist whether the IVs are continuous or dichotomized, it is typically much easier to perform simple effects tests and graph the cell means than to examine the interaction continuously. Individual cell means are fairly easy to compute and are automatically provided as part of the ANOVA output by most statistical software packages. Simple slopes are not typically provided following a regression analysis, and instead must be computed by substituting the appropriate values into the estimated regression equation. It is also easier to plot interactions analyzed with ANOVAs than those analyzed with regression. Cell means can be easily entered into most presentation software packages to generate either bar or line graphs illustrating an interaction effect. Very few statistical packages generate graphs based on either regression equations or simple slope coefficients, so the specific points on the simple slopes plot must be estimated individually and then used to create the graph in a separate program. In addition to the difficulties involved in presenting interactions involving continuous IVs, many researchers are not familiar with tests or plots of simple slopes. This may cause some members of the audience to ignore simple slopes plots or interpret those cases incorrectly.

Although the major software packages do not currently provide ways of automatically generating simple slopes plots, such options will likely become available in the near future, given the criticisms surrounding the use of dichotomized measures. In the meantime, a number of resources are available on the Internet to help researchers interpret interactions involving continuous IVs. The following websites contain tools that will help researchers create simple slopes plots:

http://www.stat-help.com/spreadsheets.html
http://www.jeremydawson.co.uk/slopes.htm
http://people.ku.edu/~preacher/interact/index.html
http://www.upa.pdx.edu/IOA/newsom/macros.htm

Reasons Related to the Prior Use of the Variable

The field has identified theoretically meaningful cutoff points on the variable being dichotomized. Whereas some measures are almost exclusively used in research settings, others are commonly used to assist real-world decision making. For example, IQ tests have been used to determine whether a child will be admitted to an accelerated learning class, Psychopathy Checklist–Youth Version (PCL-YV; Forth, Kosson, & Hare, 2003) scores have been used to determine whether an adolescent offender will be treated in juvenile court or sent to adult criminal court, and Beck Depression Inventory (BDI; Beck, 1978) scores have been used to determine whether a patient will be referred for psychological treatment. To make these decisions easier, psychologists have commonly established threshold scores to distinguish meaningful groups. IQ scores are therefore used to identify children who are "gifted," PCL scores are used to identify offenders who are "psychopaths," and BDI scores are used to identify patients who are "clinically depressed." These classifications are then used as a basis for decision making.

The logic behind this reason for dichotomization is that researchers should treat a measure as a categorical variable in their analyses if the measure is most commonly used as a categorical variable in practice. These researchers question the extent to which research in which continuous versions of a scale have been used applies to the performance of those measures when they are dichotomized. They suggest that measures with theoretically meaningful cutoff points should be analyzed categorically because this best parallels the way the measures will be used in the field. This argument is valid if the purpose of the research is to identify how well the measure will perform in a real-world setting. However, it is not valid if the purpose of the research is to understand the relations between the construct underlying the measure and other variables or to establish the validity of the measure. In these latter cases, the study methods should focus on accurately manipulating and measuring the involved constructs and not on replicating the real-world environment (Mook, 1983). As reviewed earlier, analyzing the measure continuously typically provides researchers with more power than analyzing it dichotomously, helping to prevent erroneously nonsignificant results. Researchers investigating the theoretical relations with the underlying construct should therefore avoid dichotomizing measures in their analyses, even if the construct is commonly treated categorically in applied settings.

Researchers have typically dichotomized the variable in the past. Sometimes the validity of the methods used in research is established through a systematic analysis. Other times, the validity is established through common usage. For example, Fisher's (1925) original suggestion that researchers choose an alpha of .05 was provided with very little justification. Despite this, it has become the expected cutoff for determining statistical significance simply because it has been used so often, to the point where researchers must explicitly justify the use of any other value. Similarly, the procedure of dichotomizing continuous variables has become the accepted and expected practice in certain research domains, to the point where reviewers may question the decision of a researcher to work with a continuous rather than a dichotomized indicator. Instead of being viewed as a strength, the fact that relations estimated with continuous indicators are typically stronger than those estimated with dichotomized indicators may be perceived as a type of "cheating," because this advantage was not used by prior researchers. There is the additional issue that changing

the analytic method could make the results from the current study less comparable to those performed in the past.

In cases where dichotomized indicators do not perform as well as continuous indicators, there are several theoretical benefits to be gained from changing a tradition of using dichotomized indicators to a tradition of using continuous indicators. First, the estimated effect sizes obtained with continuous indicators would more accurately reflect the relations between the underlying constructs than those obtained with dichotomized indicators. This difference can also be important when researchers try to compare the strength of the effect of interest to those found in other literatures. Second, the actual estimated size of an effect will typically be larger when using continuous indicators. Stronger effects are commonly seen as more interesting and important, so using continuous indicators can make it easier to argue for the value of studying a particular phenomenon. Finally, since analyses performed with continuous indicators typically produce larger effect sizes, it would be easier to detect the presence of moderators and mediators of the effect of interest. Weaker effects by definition have more random error in them, making it more difficult to precisely examine how the effect varies across different levels of a moderator. Mediation effects are calculated as the product of the relation of the IV with the mediator with the relation of the mediator with the DV. Dichotomized indicators typically have weaker relations with the mediators, thereby making it more difficult to obtain significant mediation tests.

While it is true that the effects obtained using continuous and dichotomous indicators are not directly comparable, there are mathematical transformations that will allow researchers to produce comparable effect sizes adjusting for the differences. After the dichotomized statistics have been corrected, results from a study can be directly compared and contrasted with the prior work in the literature. If there is a correlation between a dichotomized indicator and a continuous outcome measure, researchers can estimate the expected value of the correlation they would have obtained using a continuous indicator using the formula

$$r\{\text{continuous}\} = r\{\text{dichotomized}\}\left(\frac{\sqrt{PQ}}{h}\right),$$

where P and Q are the proportions of observations falling into the two categories of the dichotomized variable (so $Q = 1 - P$), and h is the height of the standard normal distribution at the point at which the probability to the left of the Z is equal to either P or Q (the heights will be the same at both points; Cohen et al., 2003). To compute h, one must first use a table of the standard normal distribution to determine the value of Z that has a p value corresponding to P (Q could also be used here in place of P). After that, h can be calculated from the formula

$$h = \frac{\exp\left(-\frac{Z^2}{2}\right)}{\sqrt{2\pi}},$$

which is the probability density function for the standard normal distribution. When researchers using dichotomized variables report t statistics comparing the two groups instead of correlations, Hays (1994) and Rosenthal (1994) noted that the corresponding correlation could be computed using the formula

$$r\{\text{dichotomized}\} = \sqrt{\frac{t^2}{t^2 + df}},$$

where t is the t statistic comparing whether the two groups are different on the outcome, and df represents the degrees of freedom for that test. After computing the dichotomized correlation using this formula, researchers can use the prior formulas to determine what the relation would have been had the variable been treated continuously. In considering a relation between two dichotomized variables, one could use the data from the 2 × 2 contingency table to calculate Chambers' r_e (Chambers, 1982) or the cosine approximation to the tetrachoric correlation (Alexander, Alliger, Carson, & Barrett, 1985). These statistics are used to estimate the correlation between the latent continuous variables in ways that are not influenced by the effects of dichotomization (Alexander et al., 1985).

Conclusions

The common belief among methodologists has been that researchers should always treat continuously measured variables in a numeric form (i.e., using a continuous real-valued number system to represent variable magnitudes) and that dichotomization leads to less powerful and less accurate statistical tests. Our results indicate that the picture is a bit more complicated than that. Monte Carlo simulations verified that analyzing a continuously measured variable in its original numeric form works well regardless of the nature of the latent variable or its relation to the outcome measure of interest. However, we additionally found that there are some circumstances in which the dichotomized indicator performed just as well or even slightly better than the continuous indicator. Dichotomized indicators always appeared to be viable when the data were subjected to extreme group analysis. When extreme group analysis was not used, dichotomized indicators appeared to perform at least as well as continuous indicators when (a) the underlying distribution of the latent variable was strongly categorical, (b) the proportion of individuals assigned to each category matched the proportions found in the latent distribution, and (c) the continuous measure to be dichotomized was highly reliable. Our simulations showed that violating any of these three criteria caused the dichotomized

indicator to perform notably worse than the original continuous indicator. The performance of the continuous indicator was consistently high regardless of the nature of the latent variable or its relation to the outcome variable. We would therefore suggest that researchers use continuous indicators whenever there are doubts as to whether these criteria have been satisfied. If the criteria are met, the continuous indicator will perform at a level that is comparable to that of the proportional split. If any of the above three criteria fail to be met, however, then using the continuous indicator will provide more accurate results.

Our simulations verified that dichotomization will reduce the influence of outliers in data analysis. However, it does so without consideration of why the observations are unusual. The optimal way to handle an outlier depends on whether the unusual value was caused by a scoring error, by the influence of an external variable, by the case being from a different population, or simply by natural variability in the data. It is better to examine each outlier and determine whether it represents appropriate data. When the outliers represent appropriate data, our simulations showed that Winsorizing the variable provided better results than dichotomizing the variable.

Researchers have commonly chosen to dichotomize continuous variables because it was easier to present the results from analyses with categorical predictors than from those with continuous predictors. Although continuous and dichotomized indicators commonly lead to the same conclusions, dichotomization still reduces effect sizes (Cohen et al., 2003), which could bias meta-analytic reviews of the literature and affect the perceived importance of the variable. It may also cause errors in power analyses, leading researchers in future studies to use sample sizes that are unnecessarily large. We considered the argument that it is easier to analyze and interpret the results of interactions when the variables are categorical than when the variables are continuous. However, the methods for testing interactions with continuous variables have been well defined and can be performed with standard statistical packages. Many researchers are more accustomed to presenting the results from categorical than from continuous analyses, and most software packages provide more options for graphing the results for categorical than for continuous predictors. However, we described a number of tools that can be used to illustrate the results of models using continuous predictors, which will more accurately illustrate the trends in the data than will graphs based on dichotomized indicators.

We considered the possibility that dichotomized indicators may be more appropriate when the field has identified theoretically meaningful cutoff points. We concluded that while this is a valid justification for dichotomization if the purpose of the research is to show the performance of a dichotomized measure, it is not a valid justification when the purpose of the research is to understand relations among theoretical constructs. Sometimes researchers choose to dichotomize a variable simply because articles investigating similar phenomena have dichotomized the variable in the past. We would suggest that unless there is evidence that the variable in question has the characteristics that allow continuous and dichotomized indicators to perform equivalently, efforts to change the tradition of dichotomizing variables to one in which variables are treated continuously will benefit the researcher and the literature as a whole.

In summary, our investigation revealed situations in which the use of dichotomization is appropriate. Specifically, we feel that it is acceptable for researchers to use dichotomized indicators in the following circumstances:

1. The study uses extreme group analysis.
2. The purpose of the research is to investigate how a dichotomized measure will perform in the field.
3. The underlying variable is naturally categorical, the observed measure has high reliability, and the relative group sizes of the dichotomized indicator match those of the underlying variable.

We believe that editors and reviewers should be willing to accept analyses in which dichotomized indicators have been used when researchers can successfully argue that their study falls into one of these three situations. However, we suggest that the use of the original continuous indicators should be preferred in most other circumstances. Even in situations in which the dichotomized indicator showed an advantage over the continuous indicator, the continuous indicator still performed almost as well. Researchers have little to lose by choosing to work with continuous indicators and, at times, a great deal to gain.

References

Aiken, L. S., & West, S. G. (1991). *Multiple regression: Testing and interpreting interactions.* Newbury Park, CA: Sage.

Alexander, R. A., Alliger, G. M., Carson, K. P., & Barrett, G. V. (1985). The empirical performance of measure of association in the 2x2 table. *Educational and Psychological Measurement, 45,* 79–87.

Alf, E. F., Jr., & Abrahams, N. M. (1975). The use of extreme groups in assessing relationships. *Psychometrika, 40,* 563–572.

American Psychological Association. (2001). *Publication manual of the American Psychological Association* (5th ed.). Washington DC: Author.

Barnett, V., & Lewis, T. (1994). *Outliers in statistical data* (3rd ed.). Chichester, England: Wiley.

Beck, A. T. (1978). *Depression inventory.* Philadelphia: Center for Cognitive Therapy.

Chambers, R. C. (1982). Correlation coefficients from 2×2 tables and from biserial data. *British Journal of Mathematical and Statistical Psychology, 35,* 216–227.

Cohen, J. (1968). Multiple regression as a general data-analytic system. *Psychological Bulletin, 70,* 426–443.

Cohen, J. (1983). The cost of dichotomization. *Applied Psychological Measurement, 7,* 247–253.

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

Cohen, J., Cohen, P., West, S. G., & Aiken, L. S. (2003). *Applied multiple regression/correlation analysis for the behavioral sciences* (3rd ed.). Mahwah, NJ: Erlbaum.

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

Fechner, G. T. (1860). *Elements of psychophysics*: *Vol. 1*(H. E. Adler, Trans.; D. H. Howes & E. G. Boring, Eds.). New York: Holt, Rinehart & Winston.

Feldt, L. S. (1961). The use of extreme groups to test for the presence of a relationship. *Psychometrika, 26,* 307–316.

Fisher, R. A. (1925). *Statistical methods for research workers.* Edinburgh, Scotland: Oliver and Boyd.

Fitzsimons, G. J. (2008). Death to dichotomization. *Journal of Consumer Research, 35,* 5–8.

Forth, A. E., Kosson, D. S., & Hare, R. D. (2003). *The Psychopathy Checklist–Youth Version (PCL-YV).* Toronto, ON, Canada: Multi-Health Systems.

Hays, W. L. (1994). *Statistics* (5th ed.). Fort Worth, TX: Harcourt Brace College.

Humphreys, L. G. (1978). Research on individual differences requires correlational analysis, not ANOVA. *Intelligence, 2,* 1–5.

Humphreys, L. G., & Fleishman, A. (1974). Pseudo-orthogonal and other analysis of variance designs involving individual-difference variables. *Journal of Educational Psychology, 66,* 464–472.

Hunter, J. E., & Schmidt, F. L. (1990). Dichotomization of continuous variables: The implications for meta-analysis. *Journal of Applied Psychology, 75,* 334–349.

Keppel, G. (1991). *Design and analysis: A researcher's handbook* (3rd ed.). Englewood Cliffs, NJ: Prentice-Hall.

Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores.* Reading, MA: Addison-Wesley.

MacCallum, R. C., Zhang, S., Preacher, K. J., & Rucker, D. D. (2002). On the practice of dichotomization of quantitative variables. *Psychological Methods, 7,* 19–40.

Matell, M. S., & Jacoby, J. (1971). Is there an optimal number of alternatives for Likert scale items? Study I: Reliability and validity. *Educational and Psychological Measurement, 31,* 657–674.

Maxwell, S. E., & Delaney, H. D. (1993). Bivariate median splits and spurious statistical significance. *Psychological Bulletin, 113,* 181–190.

Maxwell, S. E., Delaney, H. D., & Dill, C. A. (1984). Another look at ANCOVA versus blocking. *Psychological Bulletin, 95,* 136–147.

Michell, J. (1997). Quantitative science and the definition of measurement in psychology. *British Journal of Psychology, 88,* 355–383.

Miethe, T. D. (1985). The validity and reliability of value measurements. *Journal of Psychology, 119,* 441–453.

Mook, D. G. (1983). In defense of external invalidity. *American Psychologist, 38,* 379–387.

Neter, J., Kutner, M. H., Nachtsheim, C. J., & Wasserman, W. (1996). *Applied linear statistical models* (4th ed.). Chicago: Irwin.

Peters, C. C., & Van Voorhis, W. R. (1940). *Statistical procedures and their mathematical bases.* New York: McGraw-Hill.

Preacher, K. J., Rucker, D. D., MacCallum, R. C., & Nicewander, W. A. (2005). Use of the extreme groups approach: A critical reexamination and new recommendations. *Psychological Methods, 10,* 178–192.

Richards, F. J. (1959). A flexible growth function for empirical use. *Journal of Experimental Botany, 10,* 290–300.

Rosenthal, R. (1994). Parametric measures of effect size. In H. Cooper & L. V. Hedges (Eds.), *The handbook of research synthesis.* New York: Russell Sage Foundation.

Ruppert, D. (1988). Trimming and Winsorization. In S. Kotz, N. L. Johnson, & C. B. Read (Eds.), *Encyclopedia of statistical sciences*: *Vol. 9* (pp. 348–353). New York: Wiley.

Schulze, R. (2004). *Meta-analysis: A comparison of approaches.* Cambridge, MA: Hogrefe & Huber.

Seber, G. A. F., & Wild, C. J. (2003). *Nonlinear regression.* Hoboken, NJ: Wiley.

Snedecor, G. W., & Cochran, W. G. (1989). *Statistical methods* (8th ed.). Ames, Iowa: Iowa State University Press.

Vargha, A., Rudas, T., Delaney, H. D., & Maxwell, S. E. (1996). Dichotomization, partial correlation, and conditional independence. *Journal of Educational and Behavioral Statistics, 21,* 264–282.

Wikman, A., & Warneryd, B. (1990). Measurement errors in survey questions: Explaining response variability. *Social Indicators Research, 22,* 199–212.

(*Appendix follows*)

Appendix

Details of the Data-Generating Model

In all the simulations, the latent variable η is generated for each data point $i = 1 \ldots N$ by obtaining a random variable $x \sim U(0,1)$ and computing

$$\eta_i = \frac{1}{1 + e^{-B(x_i - M)}}, \tag{A1}$$

where B is the continuity parameter and M is the skewness parameter. Let us denote the latent variable mean and standard deviation as $\bar{\eta}$ and s_η, respectively. Using these, we can create a standardized latent variable using the equation

$$\eta_i^* = (\eta_i - \bar{\eta})/s_\eta. \tag{A2}$$

We computed the observed variable using the equation

$$m_i = \left(\frac{R}{\sqrt{1 - R^2}}\right)\eta_i^* + e_i \tag{A3}$$

with $e \sim N(0,1)$ and R is the reliability parameter, implying that the observed variable shares R^2 variance with the latent variable (so-called true variance).

The median-split indicator c is obtained by ranking the values m_i and defining

$$c_i = \begin{cases} 0 & r(m_i) \geq N/2 \\ 1 & r(m_i) < N/2 \end{cases}, \tag{A4}$$

where $r(\cdot)$ is the rank operator and N is the total sample size. The proportional split indicator is defined in a similar way using the function

$$c_i' = \begin{cases} 0 & r(m_i) \geq M \cdot N \\ 1 & r(m_i) < M \cdot N \end{cases}, \tag{A5}$$

where M is the skewness parameter. For the simulations with outliers, we defined a parameter F and a uniform variable $p \sim U(0,1)$. We then generated the variable m as in (A3). The actual variable used in the simulations was determined using the equation

$$m_i' = \begin{cases} m_i & p_i > .05 \\ m_i + F\sigma_m & p_i \leq .05 \end{cases}, \tag{A6}$$

where F is the outlier extremity parameter. Median split and proportional split indicators are computed as in (A4) and (A5). The Winsorized indicator w is determined using the equation

$$w_i = \begin{cases} m_i & m_i < 2s_{m'} \\ 2s_{m'} & m_i \geq 2s_{m'} \end{cases}. \tag{A7}$$

For the simulations concerning the outcome variable, we wish to obtain a random variable with a general logistic relation with the latent variable, allowing the relation to be either linear or nonlinear. To achieve this, we generated a random variable φ using the equation

$$\varphi_i = \frac{1}{1 + e^{-L\left(\frac{\eta_i}{s_\eta} - .5\right)}} \tag{A8}$$

to represent the part of the outcome variable that is related to the latent variable η. φ is a function of both the original latent variable η and the linearity coefficient L. As illustrated in Figure 7, increasing the value of L makes the relation between η and φ less linear and more like a step function. Once we determine φ, we combine it with random error to produce the outcome variable y using the equation

$$y_i = \varphi_i + Qe_i, \tag{A9}$$

where $e \sim N(0,1)$. The parameter Q is chosen such that the correlation $r_{y\eta}$ between the latent variable and the outcome remains constant while we change the linearity of the relation between the two variables (i.e., as we change L). This is necessary to allow the simulations to be comparable under different levels of linearity. Formally, we can obtain $r_{y\eta} = K$ (in the simulations $K = .70$) by setting

$$Q = \frac{\sqrt{r_{\varphi\eta}^2 - K^2}}{K}. \tag{A10}$$

It is easy to verify that if the variances of φ and η are both equal to 1 and the error term e has zero correlation with both φ and η, the correlation between y and η is equal to K.

Received November 26, 2007
Revision received May 4, 2009
Accepted May 15, 2009 ■