Research in Nursing & Health, 2005, 28, 496–503

Focus on Research Methods

Why Carve Up Your Continuous Data?

Steven V. Owen,* Robin D. Froman†

University of Texas Health Science Center at San Antonio, San Antonio, Texas
Accepted 26 August 2005

Abstract: Continuous data are commonplace in social, biophysical, and health research. For various reasons, researchers often carve up data into ordered chunks. Such data carving results in less information being carried by the data, a reduction or spurious increase in statistical power, and resultant Type I or Type II errors. We give examples of data carving in selected nursing literature, and illustrate how unnecessary categorization can produce erroneous statistical results. Finally, we propose credible alternatives to data carving. © 2005 Wiley Periodicals, Inc. Res Nurs Health 28:496–503, 2005

Keywords: continuous variables; dichotomizing data

Nurse researchers often collect quantitative data for their empirical studies. Nearly all self-report surveys or questionnaires can develop continuous or discrete scores, as can most biological or physiological measurements. For descriptions of their samples, analysts typically collect biographic data, some of which may be continuous, such as age, educational level, years of experience, and so on. Quantitative data encompass ordinal measurements (e.g., rank ordered preferences; state-by-state ranking of incidence of premature birth), equal appearing interval measurements (e.g., Likert-type scales; NCLEX scores), and ratio measurements (e.g., age; hours per week on the Internet). But having collected such data, researchers sometimes reduce the precision of measurement by creating ordered categories of data, which we call *carving* the data. Data carving refers to partitioning an entire scale of scores into far fewer—often only two-categories.

The meaning of a variable should help to steer its measurement. For example, body mass index (BMI) is conceptualized as continuous in nature, and its formula (kilograms weight/meters height^2) reflects this. However, researchers and practitioners often reduce BMI to a crude ordinal scale. The national Centers for Disease Control and Prevention (2005) promote such reductionism with their carving of BMI: under 18.5 = underweight, 18.5–24.9 = normal, 25.0–29.9 = overweight, and 30.0+ = obese. Researchers sometimes further simplify BMI. Lee and Paffenbarger (2000) split BMI into five unequal levels for some of their analyses, and further coarsened it to two levels representing *overweight* and *not overweight*. Such data reduction violates the obvious: People are not merely big or not big. As we will see, this sacrifice of numbers produces imprecision in the meaning of the construct, and in measurement and analysis.

Correspondence to Steven V. Owen, Department of Pediatrics, School of Medicine, and Center for Epidemiology & Biostatistics, University of Texas Health Science Center at San Antonio, San Antonio, TX.

*Professor.

†Professor and Dean.

Published online in Wiley InterScience (www.interscience.wiley.com)
DOI: 10.1002/nur.20107

There is no particular reason or requirement that quantitative data show normality in the distribution of scores. In fact, Nunnally and Bernstein (1994) asserted that ''continuity [of scores] rather than normality is usually the more important'' (p. 116). Parametric analyses with large samples have long been known to be robust in the face of violations of assumptions, such as nonnormality. That does not mean that researchers can avoid screening and cleaning data. Improving the shape of a skewed distribution—say, by re-expressing data with a square root transformation—can often improve the statistical power of an analysis and result in more honest Type I error rates (Rasmussen, 1989; Wilkinson, Blank, & Gruber, 1996). This is especially true with a small sample (Tabachnick & Fidell, 2001). Some analysts recommend shifting to nonparametric procedures when continuous data are nonnormal, but there is little evidence that that is a better route than using parametric methods with transformed data (Rasmussen & Dunlap, 1991).

WHAT MOTIVATES DATA CARVING?

Perhaps the most common form of data carving is the median split: Scores are rank ordered, and scores on one side of the median are grouped into a category called, for example, *high*. Scores on the other side are collapsed into a single value representing *low*. This dichotomization assures approximately equal numbers of scores in each group. (Interestingly, we rarely see descriptions of what researchers did with scores exactly at the median.)

An interval scale might be split into three equal size chunks, now termed *high*, *middle*, and *low*. Or, if the analyst prefers to size categories more like a normal distribution, she might use a standard deviation criterion. Any score at or beyond $+1$ *SD* goes to the *high* group; at or below -1 *SD* goes to the *low* group; and all scores in between (about 68% of them) are assigned to the *middle* group. Yet another popular carving method is to create a three-way split (by either of the two methods above), and then toss out the middle group. The remaining two groups are thus thought to be sharply contrasting groups; there is no middle group to muddle results. However, because the middlemost group, now deleted, contains the most ''representative'' scores of the distribution, generalization to other groups or populations is more difficult. Finally, we wonder about the ethics and wastefulness of collecting data with no intention of using it.

Researchers have several quite different reasons for carving their data. The first—and possible the most legitimate—occurs in the analysis of covariance (ANCOVA). When the ANCOVA assumption of homogeneity of slope is violated, one possible repair is to shift the status of an offending covariate to a categorized ANOVA factor. For example, a researcher designs a study with a plan for a one-way ANCOVA. Participants are randomly assigned to a weight loss intervention group or to a control group. Baseline measurements on the outcome variable (weight) will form the covariate, and the dependent variable will be final weight measurements after the completion of the intervention. In screening the data, she finds that the two groups' regression slopes (final weight regressed on initial weight) are not equal. She decides to convert the covariate to an ANOVA factor: Initial weight is categorized into three levels (*high*, *middle*, and *low*), and the resulting design is a 2 (group) $\times$ 3 (initial weight) ANOVA. Now she fully expects that the ANOVA interaction term will show significance, because that is what the heterogeneity of slope problem implies. But because the original continuous measurement is shrunken to a three-level factor, there is some risk a loss of statistical power and thus, a Type II error.

A second reason for data carving occurs when researchers assume that an ANOVA model is best suited to their research question(s). So although a wide range of scores exists in the data, the data are categorized simply to create ANOVA factors. For example, a researcher wants to study the association among gender, self-esteem, and depression. She has in mind a two-way ANOVA, and arranges the model so that gender is one factor and self-esteem (grouped into *high* and *low*) is the other factor. The dependent variable is depression. You might suspect that the information lost from converting self-esteem scores into two categories results in less statistical power. But as we shall see, Type II errors may or may not result from this maneuver.

A third reason is a special case of the second. Here, data are categorized in preparing for data collection. Perhaps because of an ANOVA mindset, researchers unnecessarily pressure interval level data into fewer categories. For example:

- Indicate your age: 20–29____ 30–39____ 40–49____ 50–59____ 60–69____
- Years as a staff nurse: 1–5____ 6–10____ 11–15____ over 15____

In the age example, a researcher may decide later that five ANOVA levels is too awkward, and so

coarsen the categories even more (e.g., 20–49 and over 50). Notice in the years as staff nurse example, the first three categories are equal in size, and the fourth is not only larger, but has an undefined ceiling. Notice, also, that these categories are no longer at the interval level of measurement. We should point out that collecting categorized data will likely create differences in group sizes that may be impossible to repair later. For example, Dallender, Nolan, Soares, Thomsen, and Arnetz (1999) collected an assortment of biographic data from respondents, including length of service. Their survey partitioned this continuous variable into four categories, and the resulting cell sizes were as follows:

	Psychiatrists	Nurses
$<$1 year	8	3
1–5 years	16	24
6–10 years	11	57
$>$10 years	39	217

The gross disproportions resulting from categorizing these data made length of service comparisons quite precarious. Perhaps Dallender et al. were aware of the issue, because their chi-square analysis comparing psychiatrists and nurses omitted the first two ''length'' categories.

In passing, we point out that the results of an investigation may vary, depending on what sort of cutpoints are chosen for the data carving. For example, there is little reason to believe that a median split would deliver the same results as a three-way split with the middle group excluded. One wonders how many discrepant results in published literature are simply artifacts from researchers employing different methods of data carving.

Carving of data may occur in the multivariate situation, where there are two or more dependent variables. For example, a one-way multivariate analysis of variance (MANOVA) might be arranged with the grouping factor as three levels of years of experience, with self-esteem and attitude as the dependent variables. Because the overall tests in discriminant function analysis (DFA) are mathematically equivalent to MANOVA, the same problems can occur from data carving. Rearranging the MANOVA just described, the analyst may create a DFA with self-esteem and attitude as predictors of three levels of years of experience. Here, if categorizing affects significance testing, it may give even worse results, because DFA delivers much more detail than the associated MANOVA. More specifically, in DFA multivariate group means (centroids) are calculated and perhaps plotted, one or more discriminant functions are revealed, classification tables are arranged, partial multivariate F-ratios for each predictor are given, pairwise (two groups at a time) multivariate F-ratios calculated, and standardized discriminant weights are produced for each successive discriminant function. So beyond the overall statistical tests, damage from data carving can reverberate through the entire collection of DFA results. That is, if a MANOVA omnibus test shows spurious significance, the DFA's partial F-ratios and standardized discriminant weights are likely to be faulty as well.

Over the past 20 years, DFA's popularity in the health professions has waned, and logistic regression models have become more common. Like DFA, the logistic model uses a categorical (dichotomous or polytomous) outcome variable. Natural dichotomies (such as *alive/dead*, *disease recurrence/not*) are plentiful in the health professions, and there is no special reason to create a false one as an outcome variable. But many researchers have no hesitation in carving up one or more predictor variables. Austin and Brunner (2004) offered examples of such contrived categories in logistic regression, and demonstrated how unnecessary data carving of a predictor variable creates faulty claims of statistical significance as a result. Exaggerated significance, however, does not automatically follow from data carving. Fraley and Speiker (2003) demonstrated the opposite effect: Important effects went undiscovered when interval level data were dichotomized.

RESEARCH SUFFERS FROM CARVING CONTINUOUS DATA

One problem from data carving is a measurement issue. A continuous measurement contains as much information as its numbers will allow. At the single item level (e.g., Question 1 in a 20-item self-efficacy scale), there is not much information to begin with. And there is little improvement to be gained from trying to increase the response format from seven or nine options to, say, 100. Individual items usually lack adequate reliability, and widening the response format gives an appearance of greater precision, but in truth does not boost the item's reliability (Nunnally & Bernstein, 1994). However, when individual items are aggregated to a total (sum or mean) scale score, the continuous score that results usually delivers far

greater precision. Now, categorizing that scale score drives precision back to the item level. Feldt (2005) made a compelling case that reducing an interval level score to two or three categories harms the score's measurement properties. We expect measurement scores to be dependable indicators of constructs, and there is little point in purposely making those indicators cruder. To put this a different way, when scores are purposely dulled by creating coarse categories, the correspondence between the scores and the construct also shrinks. To return to a BMI example, body mass is supposed to be an indicator of the construct *fatness*. Partitioning the continuous BMI score into *normal* versus *overweight* not only discards statistical information about individual differences, but also alters the meaning of the construct, as though *fatness* were a simple dichotomy instead of a gradient. That is, continuous BMI and dichotomized BMI appear to be indicators of related, but not identical, constructs.

Consider the logic of reduced precision in measurement. With a median split, a person's score barely above the median is now put into the same category (*high*) as a person three standard deviations above mean. That does not make good sense or good use of the data. A companion problem is just as troubling: The person right above the median is now categorically different from an adjacent person immediately below the median. A researcher keen on carving data might reply, "Well, create more precision by making more categories!" That is an interesting point, but ponder the sample size requirements and post hoc probing of an ANOVA factor with 10 levels.

A Type II error refers to overlooking a relationship in your data that probably occurs in the population. Claiming to find something that probably does *not* occur in the population is a Type I error. Type I errors can result from excessive statistical precision, which would not usually be the case when data carving occurs. In bivariate analyses, such as simple correlation, or the one-way ANOVA, the loss of information and precision from carving a measurement will ordinarily take a Type II toll. McClelland (2002) has a clever interactive model that demonstrates this with a 1-predictor regression model. When the predictor variable is dichotomized, the r^2 shrinks from .27 to .17, and the *p*-value drifts to nonsignificance, from .02 to .07. MacCallum, Zhang, Preacher, and Rucker (2002) extended the example by dichotomizing *both* the predictor and the criterion variable. In their data, the original r^2 of .09 nearly vanished to .0016. However, they also demonstrated that effect sizes can occasionally increase because of dichotomization. Cohen (1983) estimated that dichotomizing a continuous measurement generally reduces its available variance between 20% and 67%; this in turn shrinks statistical power as though one were discarding between 20% and 67% of the original sample!

In the multivariable model, such as a factorial ANOVA, the form of error is less predictable. When two (or more) ANOVA factors are created from categorized continuous variables, unequal cell sizes and unequal variances are likely. Unequal variances is a violation of an ANOVA assumption, and although a large sample may help to overcome the violation, unequal cell sizes are harder to ignore. Unequal equal cell sizes create what statisticians call a *non-orthogonal* design. In plain talk, this means that the ANOVA factors are correlated, and that fact makes it harder to assign unique effects to each factor. When factors are correlated, disruption in one factor can ripple through the entire ANOVA model. Although they focused on dichotomized variables, Maxwell and Delaney (1993) made a convincing argument that *either* Type I or Type II errors can result from data carving for a factorial ANOVA design. Unwary researchers who think that larger samples offer some protection will find just the opposite: The Type I error rate worsens as sample size increases. And, Maxwell and Delaney remarked, when two continuous variables are categorized to suit an ANOVA, bias worsens as the correlation increases between the continuous variables. Finally, because factorial ANOVAs contain one or more interaction terms, the analyst bent on categorizing should be aware that the process grossly reduces the statistical power of any interaction tests (Cohen, Cohen, West, & Aiken, 2003).

With attendant problems in measurement precision and statistical inference, data carving eventually leads to a more general corruption of research. Focusing on the time-honored practice of dichotomizing depression scores, Hankin, Fraley, Lahey and Waldman (2005) showed the empirical harm of the practice. With a large sample of youth aged 9–17, they demonstrated convincingly that depression scores wandered from their supposed meaning when they were dichotomized. Hankin et al. argued that the entire field of depression studies is likely to suffer as long as researchers continue to dichotomize depression scores into simple but misleading categories.

Pedhazur and Schmelkin (1991, p. 539) observed a conceptual problem associated with data carving. Categorizing continuous data in order to fit an ANOVA or MANOVA model can create an unjustified perception of experimental control,

with associated causal inference. For example, Pridham et al. (2005) categorized mothers, somewhat confusingly, as "Poverty = 1, status of *at or above* poverty level = 0" (emphasis added; p. 262). In their quasi-experimental study, Pridham et al. repeatedly alluded to the causal *effect* of poverty on mothers' feeding competencies. Pridham et al.'s study was generally well designed, but the causal inference goes beyond slippery semantics. Rosenthal (1994) has termed inappropriate causal language *causism*, and defined it as an important ethical issue in the conduct and reporting of research. "If a perpetrator of causism is unaware of the causism," he wrote, "its presence simply reflects poor scientific training. If the perpetrator is aware of the causism, it reflects blatantly unethical misrepresentation and deception" (p. 128).

EXAMPLES OF DATA CARVING IN NURSING RESEARCH

How commonplace is data carving in nursing research? For a partial answer, we studied all quantitative reports from each issue of *Research in Nursing & Health* over a recent 1-year period (six issues from August 2004 through June 2005). The issues contained between five and eight quantitative studies. Of a total of 40 articles, 16 (40%) contained some form of data carving.[1] It may be that data carving occurs more frequently in nursing research than in other areas of the health professions, or in other psychosocial fields. MacCallum et al. (2002) reviewed over 900 articles in three psychology journals, and found that over 11% of them did unnecessary categorizing of data. In any case, data carving is not a rare event.

We now turn to several examples of data carving in nursing research. Cimprich (1998) compared attentional fatigue of patients who received breast conservation surgery versus those who received mastectomy. In one three-way ANOVA model, age was chopped into three levels: *younger* (age 25–45, $n = 14$), *middle* (age 46–64, $n = 44$), and *older* (age 65–79, $n = 16$). Here, categorizing created wide differences in cell size, which in turn caused the age effect to be correlated with the other two effects (and their interactions) in the ANOVA model. In passing, we note that Cimprich's use of ANOVA with non-experimental data may have encouraged her to use causal language throughout her report.

Nies, Buffington, Cowan, and Hepworth (1998) repeated a race (white vs. black) X body size (obese vs. nonobese) ANCOVA seven times for each of the subscales of the Health-Promoting Lifestyle Profile. They categorized body size via a three-way split on BMI, and excluded the middle group. In these analyses, subgroup sizes were approximately equal. Nies et al. also took the liberty of using causal language with their nonexperimental data, confluding, for example, that "obesity . . . influences health promoting behaviors" (1998, p. 255). With their data, however, the opposite hypothesis is just as plausible: Health promoting behaviors influence obesity.

White and Frasure-Smith (1995) compared perceived stress and uncertainty of coronary angioplasty patients versus patients who had bypass surgery. In their various ANOVAs, one categorized factor was perceived social support, which was partitioned at the mean to create *high* and *low* groupings. They did not reveal whether the categorization resulted in unequal group sizes. In their discussion of findings, White and Frasure-Smith properly avoided causal language, acknowledging "[a] major weakness of descriptive, comparative research of this kind is . . . the consequent inability to draw cause and effect conclusions" (p. 24).

Dowe, Lawrence, Carleson, and Keyserling (1997) studied patients' learning from drug literature written at varying readability levels. For a two-way ANOVA, they split educational background into three levels: *completed grades 0–8* ($n = 40$), *grades 9–11* ($n = 51$), and *high school -+* ($n = 79$). Although the three written passages' readability levels had been randomly assigned to participants, categorizing the educational variable resulted in highly disproportionate cell sizes, ranging from 8 to 23. In this quasi-experimental study, Dowe et al. were ambivalent about causal inference, sometimes avoiding it (variables were "associated" [p. 91]) and sometimes using it ("influence" [p. 92]).

EXAMPLE OF HOW DATA CARVING CAN ALTER STATISTICAL RESULTS

Using data from an unpublished study, we compare results from an ANOVA using a categorized variable, and results from the same data, but without the data carving, using regression analysis. The ANOVA is a 2×2 model, with the factors

[1]Reference list of articles containing data carving available from Owen.

Table 1. Summary of 2 × 2 ANOVA With Categorized Variable Anxiety

Source	*SS*	*df*	*MS*	*F*	*p*	η^{2*}
Gender	0.103	1	0.103	3.23	.073	.013
Anxiety	0.748	1	0.748	23.47	$<.001$	.093
G × A	0.208	1	0.208	6.50	.007	.026
Error	7.012	220	0.032			

*η^2 (eta squared) is an ANOVA effect size, representing the proportion of variance in the dependent variable explained by grouping status. From Cohen's (1988) rough guidelines, anxiety's .09 represents a medium effect.

anxiety and gender. Anxiety had been measured with a 20 cm visual analog scale, and was carved by a median split into *high* and *low* categories. The dependent variable was perceived pain. Table 1 summarizes the ANOVA. The large sample ($N = 224$) creates abundant statistical power, so we also show effect sizes (η^2), which are calculated independently of sample size. The ANOVA main effect for anxiety was significant, and had an effect size of .09 (i.e., anxiety grouping status explained 9% of the variation in perceived pain). However, that main effect is overridden by the significant interaction term, which shows a smaller effect size of .03.

Next, the data were re-run as a regression model, in which anxiety is left as a continuous variable. The regression summary, in Table 2, shows somewhat different results. As before, anxiety is significant, but now it stands on its own, because the interaction term is nowhere close to significant. Thus, two of the three tested effects are disrupted by data carving. The categorized anxiety appeared to create a Type I error by assigning spurious significance to the interaction term. The larger message from this example is that data carving created untrustworthy significance tests, which in turn delivered a misleading story about these data.

RECOMMENDATIONS

Problems associated with data carving continuous measurements have been known for over 30 years, and methodologists have frequently issued warnings about the procedure. Yet the practice continues in mainstream scholarship. We speculate that there are two main reasons for the practice. First, nurse researchers (including journal reviewers, and editors) may not have had adequate training in research design and statistics. Most introductory statistics instructors and their texts, after all, say nothing whatsoever against carving continuous data into categories. And even more advanced texts imply that data carving is acceptable. For example, in their widely used multivariate book, Tabachnick and Fidell (2001) give a MANOVA illustration where one factor is (learning) disability, carved into levels of *mild*, *moderate*, and *severe* (pp. 332–351). Second, principles of observational learning (Bandura, 1986) suggest that the published literature, which contains so many examples of the practice, tacitly legitimizes data carving and encourages others to do so. Social diffusion of a questionable practice can be hard to correct.

Our main recommendation is to let continuous data be continuous. This means that you should not force fit an ANOVA design on certain data. It also may mean that researchers will have to grapple with the idea that ANOVA is just a special—and more restrictive—case of multiple regression (see, for example, Kleinbaum, Kupper, Muller, & Nizam, 1998; Pedhazur, 1997). And, of course, it means that some of us will have to set aside our familiar and well fitting ANOVA thinking cap and try on a regression model for size. In the end, though, we should discover that

Table 2. Summary of Regression With Continuous Variable Anxiety

Predictor	*b*	se_b	*Beta*	*t*	*p*	sr^{2*}
Gender	−.005	.017	−.011	−0.29	.76	.033
Anxiety	.080	.019	.192	4.20	$<.001$	.031
G × A	.013	.016	.034	0.82	.44	.001

*sr^2 (squared semipartial correlation) is a multiple regression effect size for individual predictors. From Cohen's (1988) rough guidelines, anxiety's .03 represents a small effect.

regression analysis helps to avoid the pitfalls of categorizing data, and in doing so, creates more credible and replicable research results that reduce measurement problems, and both Type I and Type II error rates.

The use of multiple regression as an all-purpose analytic system has been advocated for some time (e.g., Cohen, 1968; Pedhazur, 1997). Surely there are instances, such as a repeated measures design, where regression is clumsier than the straightforward ANOVA approach. But the analyst willing to consider regression as a useful alternative to ANOVA has two important advantages. First, it removes the opportunity for damaging data by carving continuous data, and second, it creates a far more flexible approach to data analysis.

In the multivariate case, where data carving is used with MANOVA or DFA, the parallel analysis for using continuous data depends on design complexity. In the one-way MANOVA and DFA example discussed earlier, the design simplifies to a multiple regression problem, where self-esteem and attitude are used to predict the continuous—not categorized—variable, years of experience. With factorial MANOVA or DFA models, the preferred approach would be a canonical correlation analysis, which is the more general case of both MANOVA and DFA. Canonical analysis, like DFA, will produce standardized coefficients, partial multivariate *F*-ratios for each variable, and possibly two or more discriminant functions. Unlike DFA, it will not give classification matrices showing how well the predictors can sort persons into their known groups. But then if categorizing continuous variables creates spurious classifications, it may be better not to have any classifications at all. Cohen et al. (2003) and Pedhazur (1997), give detailed information about the symmetry among MANOVA, DFA, and canonical analysis, and they also support our argument: Why carve up your data?

We have taken a strong position about avoiding categorizing continuous data, and are not the first to do so. Humphries (1978) used especially forceful language, calling data carving ''crude, misleading'' (p. 874). Cohen (1983) also argued against it, and more recently, MacCallum et al. (2002) reinforced the case. But like all rules of thumb, there are important exceptions. We argued that Type I or Type II errors *can* occur from data carving, but in truth, little is known about the prevalence or magnitude of such errors in real data sets. It may be that 90% of the studies that contained data carving would deliver substantially the same conclusions if the data had been treated as continuous. Also, in the health sciences, many naturally continuous variables are rendered discrete for very practical purposes. For example, despite ranges of several symptoms, people typically are classified as either having a disease or not. An intervention is given or it is not. Surgery is performed or avoided. A person is diagnosed as hearing impaired or not. Courses of action are thus simplified to dichotomous decisions; the world would be far too complicated to have as many levels of decisions as levels of the original measurements. Thus, we will temper the ''Why carve up your data?'' with a gentler request: Please *resist* carving your data.

REFERENCES

Austin, P.C., & Brunner, L.J. (2004). Inflation of the type I error rate when a continuous confounding variable is categorized in logistic regression analyses. Statistics in Medicine, 23, 1159–1178.

Bandura, A. (1986). Social foundations of thought and action. Englewood Cliffs, NJ: Prentice-Hall.

Centers for Disease Control and Prevention. (2005). BMI—Body mass index: BMI calculator. Retrieved June 22, 2005 from http://www.cdc.gov/nccdphp/dnpa/bmi/calc-bmi.htm.

Cimprich, B. (1998). Age and extent of surgery affect attention in women treated for breast cancer. Research in Nursing & Health, 21, 229–238.

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, 249–253.

Cohen, J. (1988). Power analysis for the behavioral sciences (2nd ed.). Hillsdale, NJ: Lawrence 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: Lawrence Erlbaum.

Dallender, J., Nolan, P., Soares, J., Thomsen, S., & Arnetz, B. (1999). A comparative study of the perceptions of British mental health nurses and psychiatrists of their work environment. Journal of Advanced Nursing, 29, 36–43.

Dowe, M.C., Lawrence, P.A., Carlson, J., & Keyserling, T.C. (1997). Patients' use of health-teaching materials at three readability levels. Applied Nursing Research, 10, 86–93.

Feldt, L.S. (2005). Estimating the reliability of dichotomous or trichotomous scores. Educational and Psychological Measurement, 65, 28–41.

Fraley, R.C., & Spieker, S.J. (2003). Are infant attachment patterns continuously or categorically distributed? A taxometric analysis of strange situation behavior. Developmental Psychology, 39, 387–404.

Hankin, B.L., Fraley, R.C., Lahey, B.B., & Waldman, I.D. (2005). Is depression best viewed as a continuum

or discrete category? A taxometric analysis of childhood and adolescent depression in a population-based sample. Journal of Abnormal Psychology, 114, 96–110.

Kleinbaum, D.G., Kupper, L.L., Muller, K.E., & Nizam, A. (1998). Applied regression analysis and other multivariable methods. Pacific Grove, CA: Duxbury Press.

Lee, I-Min, & Paffenbarger, R.S., Jr. (2000). Associations of light, moderate, and vigorous intensity physical activity with longevity. American Journal of Epidemiology, 151, 293–299.

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.

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

McClelland, G. (2002). Negative consequences of dichotomizing continuous predictor variables. Retrieved 22 June 2005 from http://psych.colorado.edu/~mcclella/MedianSplit/.

Nies, M.A., Buffington, C., Cowan, G., & Hepworth, J.T. (1998). Comparison of lifestyles among obese and nonobese African American and European American women in the community. Nursing Research, 47, 251–257.

Nunnally, J.C., & Bernstein, I.H. (1994). Psychometric theory (3rd ed.). New York: McGraw-Hill.

Pedhazur, E.J. (1997). Multiple regression in behavioral research (3rd ed.). New York: Harcourt Brace College Publishers.

Pedhazur, E.J., & Schmelkin, L.P. (1991). Measurement, design, and analysis: An integrated approach. Hillsdale, NJ: Lawrence Erlbaum.

Pridham, K, Brown, R, Clark, R., Limbo, R.K., Schroeder, M., Henriques, J., et al. (2005). Effect of guided participation on feeding competencies of mothers and their premature infants. Research in Nursing & Health, 28, 252–267.

Rasmussen, J.L. (1989). Data transformation, Type I error rate, and power. British Journal of Mathematical and Statistical Psychology. 42, 203–213.

Rasmussen, J.L., & Dunlap, W.P. (1991). Dealing with nonnormal data: Parametric analysis of transformed data vs. nonparametric analysis. Educational and Psychological Measurement, 51, 809–820.

Rosenthal, R. (1994). Science and ethics in conducting, analyzing, and reporting psychological research. Psychological Science, 5, 127–133.

Tabachnick, B.G., & Fidell, L.S. (2001). Using multivariate statistics (4th ed.). Boston: Allyn and Bacon.

White, R.E., & Frasure-Smith, N. (1995). Uncertainty and psychologic stress after coronary angioplasty and coronary bypass surgery. Heart & Lung: Journal of Critical Care, 24, 19–27.

Wilkinson, L., Blank, G., & Gruber, C. (1996). Desktop data analysis with SYSTAT. Upper Saddle River, NJ: Prentice Hall.