MIS Quarterly

RESEARCH ESSAY

IN PURSUIT OF MODERATION: NINE COMMON ERRORS AND THEIR SOLUTIONS[1]

By: **Traci A. Carte**
Division of Management Information Systems
Price College of Business
University of Oklahoma
307 W. Brooks Drive
Norman, OK 73019-0450
U.S.A.
tcarte@ou.edu

Craig J. Russell
Division of Management Information Systems
Price College of Business
University of Oklahoma
307 W. Brooks Drive
Norman, OK 73019-0450
U.S.A.
cruss@ou.edu

Abstract

One result of the increasing sophistication and complexity of MIS theory and research is the number of studies hypothesizing and testing for moderation effects. A review of the MIS and broader management literatures suggests researchers investigating moderated relationships often commit one or more errors falling into three broad categories: inappropriate use or interpretation of statistics, misalignment of research design with phenomena of interest, and measurement or scaling issues. Examples of nine common errors are presented. Commission of these errors is expected to yield literatures characterized by mixed results at best, and thoroughly erroneous results at worse. Procedures representing examples of best practice and reporting guidelines are provided to help MIS investigators avoid or minimize these errors.

Keywords: Tests of moderation, contingency models, PLS

ISRL Categories: AI0604; AI0610; AI0702; IB01

Introduction

Lee (2001) argued that the contribution many university researchers make to the MIS field is "scrupulous attention" to scientific methods, using largely quantitatively and statistically based approaches. MIS researchers have recently focused on improving the quantitative methods employed.

[1]V. Sambamurthy was the accepting senior editor for this paper.

For example, MIS researchers investigated methodological issues in experiments (Jarvenpaa et al. 1985), highlighted problems of statistical power (Baroudi and Orlikowski 1989), questioned model complexity (Lee et al. 1997), and examined the rigor with which instruments are validated (Boudreau et al. 2001). The goal of this paper is to sensitize MIS researchers to methodological issues surrounding tests of moderated relationships.

Three types of relationships dominate MIS research: simple linear or additive relationships, mediated relationships (typically sequences of linear relationships), and moderator relationships. Moderator relationships are the most interesting and perhaps the most difficult of the three to establish empirically (McClelland and Judd 1993). A review of recent MIS research reveals an increasing interest in moderated relationships. From 1991 through 2000, *MIS Quarterly*, *Information Systems Research,* and *Journal of Management Information Systems* published 26 articles directly testing moderated relationships (see Appendix A). *MIS Quarterly* and *Information Systems Research* had 17 articles suggesting but not testing moderation in the same 10-year period.

The increasing interest in moderated relationships reinforces a notion that MIS researchers are increasingly addressing: context matters in MIS research. Relevant contexts include organizational, technological, and individual. For example, researchers investigating technology acceptance have incorporated individual contexts such as personal innovativeness (Agarwal and Prasad 1998), work experience and gender (Venkatesh and Morris 2000) and yielded a richer understanding of the phenomenon of interest.

This paper critically assesses moderation tests performed by MIS researchers. We hope to raise awareness about common errors and enhance the craftsmanship of moderation testing by providing a central summary of nine common errors. While these errors have been separately identified elsewhere, this is the first attempt to synthesize and assess the extent to which MIS researchers are prone to their commission. Some of these errors, while generally understood, still occur frequently. Others are less well understood and occur with great regularity. Importantly, the increasingly popular use of partial least squares (PLS) applications (Gefen et al. 2000) has been accompanied by an introduction of a new error as well as reintroduction of some old errors.

We critically assess moderation tests in the sample of 26 articles published from 1991 through 2000, identifying three general types of errors labeled inappropriate statistics, misalignment of phenomena and research design, and measurement issues. Nine specific errors were distinguished, although not all studies reported enough information to determine whether an error occurred. Descriptions of these errors and methods of avoiding them should help MIS investigators advance theory and practice by minimizing Type I and Type II errors in tests of moderation

We first review various conceptual definitions of moderation, then present three sets of common difficulties encountered when searching for moderation in MIS research and ways to avoid them. Analysis of select articles is presented to demonstrate error commission, potential consequences, and illustrations of best research practice. We conclude by recommending reporting guidelines to improve the thoroughness with which authors report moderation-related evidence and enhance the ability of readers and reviewers to evaluate tests of moderation.

Definitions

Review of moderation definitions revealed what at first appeared to be an unsettlingly high level of variation. Fortunately, evidence supporting the presence of virtually all conceptualizations of moderation in applied behavioral field research can be assessed using hierarchical moderated multiple regression (MMR, Saunders 1956) to test H_0: $\Delta R^2 = R^2_{mult} - R^2_{add} = 0$ using least squares procedures (ordinary or PLS), where:

$$\hat{Y} = b_0 + b_1X + b_2Z;\ R^2_{add} \qquad \textbf{Equation 1}$$

$$\hat{Y} = b_0 + b_1X + b_2Z + b_2XZ; \quad R^2_{mult} \qquad \textbf{Equation 2}$$

An F statistic derived using Equation 3 that is significantly greater than 1.00 leads to rejection of H_0: $\Delta R^2 = 0$ and the conclusion that either Z moderates the X→Y relationship or X moderates the Z→Y relationship.[2] Using this procedure, large values of ΔR^2 occur when any one of a number of conceptualizations of moderation occurs.

$$F_{1,N-3} = \frac{\Delta R^2}{(1 - R^2_{mult}) / (N - 3)} \qquad \textbf{Equation 3}$$

Definitions of moderation provided in the literature are summarized in Table 1. Of particular note is Arnold's (1982, 1984, amplified by Baron and Kenney 1986) distinction between circumstances where the strength of the X→Y relationship varies as a function of Z versus the nature of the X→Y relationship varies as a function of Z. The former is often referred to as differential validity while the latter is referred to as differential prediction.[3] The distinction between these two types is important as differential prediction is the form of moderation appropriately tested for using MMR. The definition of moderation applied in this study is that of differential prediction, where the nature of the X→Y relationship varies as a function of Z.

MIS researchers are not consistent in their moderation conceptualizations. For example, a number of MIS investigators incorrectly use differential validity and differential prediction interchangeably. Four articles in our sample included language describing moderation as differences in strength of the X→Y relationship *and* differences in the nature of the X→Y relationship (Devaraj and Kohli 2000; Hardgrave et al. 1999; Harrison et al. 1997; McKeen et al. 1994). By way of illustration, McKeen et al. (1994) stated they examined whether "the strength of the participation-satisfaction relationship depended on the level of" (p. 427) task complexity and other moderators. However, these authors did not report differences in strength of participation-satisfaction (i.e., $r_{participation\text{-}satisfaction}$) across levels of task complexity, instead reporting differences in the nature or slope of the participation-satisfaction relationship across levels of task complexity.

Importantly, insight into underlying processes behind moderation is most likely to result from qualitative research efforts aimed at adding meaning to abstract relationships found in quantitative research. Such efforts will be most justified when empirical evidence suggests the presence of an underlying moderation process. In field studies using random effects designs (by far the dominant research design used in applied behavior research), MMR procedures and recent PLS variants constitute the dominant method of detecting moderation effects (Aiken and West 1991). The nine common errors discussed below address interpretations of MMR and PLS results used to test the definition of moderation described above.

Nine Common Errors

Unfortunately, even a casual reader of research in MIS, organizational behavior, human resources management, organizational theory, and strategy can find examples of ill-advised or outright inappropriate research methods in studies examining moderation effects. Examination of the MIS research generated a list of nine common errors that cause severe problems. These are summarized in Table 2 and grouped into three categories based on our views of underlying similarities: (1) inappropriate use or interpretation of statistics, (2) misalignment of phenomena and research design, and (3) measurement or scaling issues.

---

[2]Note that mathematically the test of H_0: $\Delta R^2 = 0$ is the same as an omnibus test of whether b_0 and b_1 for the following two equations are significantly different from one another:

$$\hat{Y} = b_1 + b_2X; \text{ for } Z = 1$$

$$\hat{Y} = b_1 + b_2X; \text{ for } Z = 2$$

[3]Interested readers may contact the second author for more information on the distinction between differential validity and differential prediction.

Table 1. Definitions of Moderation

Citation	Definition of Moderation
Jaccard, Turrisi, and Wan (1990)	Moderation occurs when the relationship between X and Y depends on Z.
Cohen and Cohen (1983)	Moderation occurs when X and Z have a joint effect in accounting for incremental variance in Y beyond that explained by X and Z main effects.
Baron and Kenney (1986)	A moderator variable is a "variable that affects the *direction* and/or *strength* of the relationship between an independent or predictor variable and a dependent or criterion variable" (p. 1174, emphasis added).
James and Brett (1984)	Z is a moderator when "the *relationship* between two (or more) other variables, say X and Y, is a function of the level of" Z (p. 310, emphasis added).
Cortina (1993)	moderation occurs when "the *effect* of one variable, X, on another variable, Y, depends on the level of some third variable," Z (p. 916, emphasis added).
Schmitt and Klimoski (1991)	"a moderator variable affects the *nature of the relationship* between two other variables" (p. 18, emphasis added)
Arnold (1982, 1984, amplified by Baron and Kenney 1986)	Offer two definitions, distinguishing between circumstances where the strength of the X→Y relationship varies as a function of Z versus the nature of the X→Y relationship varies as a function of Z. The former is often referred to as differential validity while the latter is referred to as differential prediction.
Sharma, Durand, and Gur-Aire 1981	Offer a slightly different perspective on differential validity versus differential prediction. They refer to differential prediction as "pure moderators" and differential validity as "homologizer variables." Homologizer variables are those that affect the criterion through the error term.

In identifying illustrations from our sample, we soon discovered that reporting standards in MIS do not routinely include enough information to assess commission of these errors. For most errors we summarize information reported that contributed to our evaluation of the likelihood an error was committed. Appendix A summarizes each article's assessment.

Inappropriate Use or Interpretation of Statistics

Solutions to problems in this first category are fairly straightforward: investigators should appropriately use and interpret statistical procedures. Examples from the literature are used to describe two problems and solutions in this category.

Error 1: Interpreting b_3 Instead of ΔR^2

Arithmetically, test statistics regarding H_0: $b_3 = 0$ and H_0: $\Delta R^2 = 0$ parallel one another and always yield the same conclusions. While this is true about the test statistics, the population parameters $\Delta\rho^2$ and β_3 are generally not parallel or equal representations of moderator effect size. In fact, ΔR^2 and b_3 are only equal when the XZ interaction is measured without error and the variance of Y (s^2_Y) is equal to the variance of the product term (s^2_{XZ}). ΔR^2 and b_3 are not generally even linearly related. Only the sample estimate ΔR^2 is a reflection of moderator effect size.

Chin et al. (1996) recently noted that,

> in addition to the change in R^2, the estimated beta for the interaction term pro-

Table 2. Nine Common Errors of Commission in Conclusions Drawn about Moderation Effects

#	Error Description	Error Solution
	Inappropriate Use or Interpretation of Statistics	
1	Using b_3 instead of ΔR^2 as an index of moderator effect size	Use ΔR^2 as the index of moderator effect size after establishing statistical significance using either a t-test of H_0: $b_3 = 0$ or H_0: $\Delta R^2 = 0$.
2	Interpreting b_1 and b_2 when X and Z are interval scale measures	Develop ratio scale measures of X and Z or do not use or develop models requiring interpretation of b_1 and b_2.
	Misalignment of Phenomena and Research Design	
3	Confounding of X•Z with X^2	Partial out X^2 effects by adding X^2 term to MMR analyses.
4	Incorrect specification of the $X \rightarrow Y$ versus $Y \rightarrow X$ causal sequence.	1. Careful consideration of theory or rationale justifying causal sequence to ensure correct sequence is selected. 2. Examine the moderation effects in both causal sequences as part of exploratory efforts that might lead to theory development.
5	Low power of random effects designs	1. Estimate sample size required to reject H_0: $\Delta R^2 = 0$ with X, Z combinations that are expected to be observed in the data. 2. Take extra care before "trimming" any outliers.
	Measurement or Scaling Issues	
6	Dependent variable scale is too coarse	Investigate number of levels of X and Z expected and select method of operationalizing Y that meets or exceeds their product.
7	Nonlinear, monotonic Y transformations	Do no transformations without a theoretical rationale. Bootstrap estimates of confidence interval around ΔR^2 if parametric assumptions are not met.
8	Influence of measurement error on X•Z.	First, estimate expected ΔR^2 by simulating X•Z interaction and adjusting obtained ΔR^2 for measurement error in X and Z. Second, estimate sample size required to reject H_0: $\Delta R^2 = 0$ when the expected MMR effect size is the adjusted estimate of ΔR^2.
9	Gamma Differences between two Groups in PLS.	Test for differences between inter-item correlation matrices between two groups using Hotelling T^2 and/or assess factor loading similarities using coefficient of concordance (Harman 1976). If no differences exist, scales derived from the items must be arrived at in the same way for all observations. If differences exist, explore for possible differences in latent construct domain tapped by items.

> vides additional information regarding the interaction effect. This estimate informs us as to how much a unit change in the moderator variable Z would change the regression relationship of Y on X. (p. 22)

Unfortunately, when X, Y, and Z are measured on interval scales, the units of measurement are arbitrary. Change in the X-Y relationship associated with a unit change in Z can be artificially inflated or deflated by simply changing Z's scale of measurement. Further, multicolinearity between X, Z, and the XZ product term causes additional b_3 distortion.[4]

After making this incorrect assertion, Chin et al. focused on b_3 estimates in reviewing 70 MIS studies reporting tests of moderation since 1980. In their Table 2 summarizing studies using regression and path analytic techniques, Chin et al. reported b_3 terms as evidence of moderator effect size and concluded that

> the literature consistently reported moderators with a small effect size, beta averaging 0.10, suggesting that moderating terms play only a small part for understanding information systems issues. (p. 23)

In fact, as b_3 is not an indicator of moderator effect size, no conclusion can be drawn about the role moderators play in understanding information systems issues. Chin et al. could have formed a conclusion about the role of moderators if they had summarized ΔR^2 across studies.

Unfortunately, Chin et al. may have been limited by the information reported in their studies and unable to draw strong conclusions about the role of moderators in MIS research. Only seven articles (27 percent) in our sample actually reported ΔR^2. In one best practice example, Harrington (1996) investigated moderating effects of denial of responsibility on codes of ethics and their relationship to computer abuse judgments and intentions. Her analysis included not only a calculation but a discussion of ΔR^2 effect size.

Solution. Investigators must use ΔR^2 to draw conclusions about relative moderator effect sizes; use of b_3 will lead to spurious conclusions.

Error 2: Interpreting b_1 and b_2 When X and Z are Interval Scale Measures

Error 2 occurs when X and Z are measured on interval scales and investigators attempt to interpret b_1 and b_2 in Equation 2. There are two potential problems with interpreting these statistics: variability due to linear transformation and/or confounding main and moderating effects.

To our knowledge, Schmidt (1973, footnote 4) first noted b_1 and b_2 could vary greatly after linear transformations of X and Z. If X and Z are measured using interval scales, the information contained in those measures remains unchanged when a constant is added to or subtracted from them or they are multiplied or divided by a constant—all linear transformations of X and Y are equally legitimate and viable. Unfortunately, b_1 and b_2 in Equation 2 do not stay the same if X and Z are subjected to such changes, although ΔR^2 and the test statistics for H_0: $\Delta R^2 = 0$ and H_0: $b_3 = 0$ are not affected.

An alternate way of describing this problem is captured by Figures 1a and 1b. Both describe path models involving an X→Y relationship that is moderated by Z. Figure 1b differs in that a direct Z→Y relationship is also hypothesized. Figure 1a describes the model $Y = b_0 + b_1X \bullet Z$ while Figure 1b describes the model $Y = b_0 + b_1X + b_2Z + b_3X \bullet Z + e$ ($b_1 = 0$ in Figure 1b). Schmidt showed it is impossible to differentiate between these models in the presence of interval scale measurement. Unfortunately, many examples of models similar to Figure 1b and interpretations of b_1 and b_2 appear in applied behavioral research.

[4]Interested readers may contact the second author for more detail on how multicolinearity affects estimates of b_3 but not tests of H_0: $\Delta R^2 = 0$.

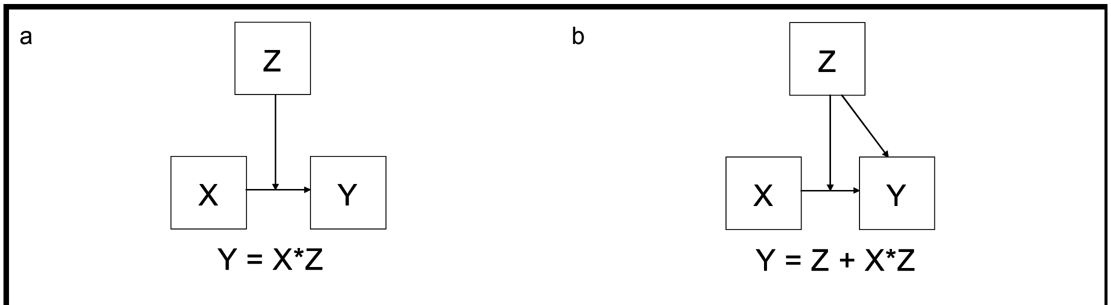

Figure 1. Two Interactive Path Models

Fifteen articles in our samples hypothesized main and moderating effects. Three of these 15 illustrate best practices by employing ratio scales (Ahituv et al. 1998; Banker and Slaughter 2000; Devaraj and Kohli 2000). In fact, concurrent main and moderating effects may be theoretically justified. Error 2 occurs as a shortcoming of traditional data analysis techniques such as regression and ANOVA. Five of the 15 articles used subgroup analysis to avoid the issue addressed here. It introduced a different concern as the main effects of these articles were interpreted from analysis that did not include interaction terms, resulting in biased estimates in an underspecified model.

In one example, Harrison et al. (1997) hypothesized that attitude, subjective norm, and perceived behavioral control would directly impact strategic decision to adopt new IT. They also hypothesized that these relationships would be moderated by organization size. Both sets of hypotheses were tested and interpreted using MMR. As Harrison et al. used interval scales, main and interaction effects could not simultaneously be examined; the main effects they report are uninterpretable.

In a best practice example, Banker and Slaughter (2000) investigated the main effect of software structure on enhancement costs, and the moderating effects of software structure of the relationships between volatility, complexity, and enhancement costs. They did not commit this error because the measures employed were ratio scaled. An additional 11 articles in our sample avoided this error by hypothesizing moderating effects only, or by interpreting main effects only after moderating effects were found to be insignificant (i.e., McKeen et al. 1994).

Solution. Unfortunately, the only way to interpret b_1 and b_2 in Equation 2 is when X and Z are measured on ratio scales. Creating ratio scale measures of organizational members' perceptions requires advanced psychophysical scaling procedures (e.g., Birnbaum 1985, 1989, 1998) and substantial pre-study scale development efforts (for an example, see Arnold 1981). When ratio scales are not available, as is the case for many important MIS phenomenon, investigators must avoid models such as those portrayed in Figure 1b and resist temptations to interpret b_1 and b_2.

Misalignment of Phenomena and Research Design

Solutions and problems in the next two categories often depend on the research goal. Steps available when testing strong theory-based moderation predictions are constrained by the theory's specifications. These constraints may contribute substantial power to investigators (e.g., Bobko 1986), although constraints can make tests of a theory virtually impossible (e.g., Podsakoff et al. 1995). This is especially true when constructs cannot be operationalized at appropriate levels of measurement. More steps are available when investigators attempt to build theory from exploratory analyses results (Glaser and Strauss 1967).

Errors in this category occur when investigators make basic research design decisions that are incongruent with questions being asked of the phenomena under investigation.

Error 3: Confounding of X•Z with X^2

Cohen (1978) demonstrated how a curvilinear relationship between X and Y is very similar to conceptualizations of moderation. If moderation occurs when the relationship of X and Y depends on the level of Z, a curvilinear X→Y association suggests the X→Y relationship depends on the level of X. In a survey of 123 significant MMR interaction effects reported in the *Journal of Applied Psychology* in 1991 and 1992, Cortina (1993) found multicolinearity (r_{xz}) ranged from 0 to .68 with an average of .21. Hence, Lubinsky and Humphrey's (1990) speculation that significant moderators may be simply nonlinear X→Y effects in disguise would seem to be a possibility in those studies with relatively high multicolinearity (r_{xz}), although not an excessively common problem.

In an MIS illustration, Igbaria et al. (1994) investigated the moderating role of job involvement on the relationships between work experiences, expectations, and attitudinal outcomes for IS personnel. Previous results suggested job involvement was quadratically related to career stage (Raelin 1985) and tenure (Wagner 1987). To the extent that job involvement is highly correlated with career expectations, $X_{JI}^2 \cong X_{JI} \bullet Z_{CE}$. Because of this, the results shown in Table 5 of Igbaria et al may inaccurately confound moderation ($X_{JI} \bullet Z_{CE}$) and nonlinear (X_{JI}^2) effects.

Only seven articles in our sample reported correlation matrices without which the likelihood of this error (in the form of high r_{xz}) cannot be determined. In articles reporting correlation matrices, the r_{xz} correlations ranged from .008 to .883 (weighted-average r = .187). The correlation of .883 (Banker and Slaughter 2000) suggests this error may have occurred. No illustration of best practices for avoiding this error was found in our sample because no author provided evidence (described below) that the error was not committed. There were, however, several articles that reported very low r_{xz} correlations (e.g., McKeen and Guimaraes [1997] and McKeen et al. [1994] reported r_{xz} ranging from .008 to .05, indicating this error was unlikely).

Solution. Cortina's proposed solution would slightly decrease Equation 3's statistical power (i.e., the F statistic of H_o: $\Delta R^2 = 0$). Decreased power would most likely be negligible as it would only involve reducing the F statistic denominator degrees of freedom by 2. The solution modifies MMR to a three-step process examining ΔR^2 for the equations 4, 5, and 6:

$$\hat{Y} = b_0 + b_1X + b_2Z \qquad \textbf{Equation 4}$$

$$\hat{Y} = b_0 + b_1X + b_2Z + b_3X \bullet X^a + b_4Z \bullet Z^c \qquad \textbf{Equation 5}$$

$$\hat{Y} = b_0 + b_1X + b_2Z + b_3X \bullet X^a + b_4Z \bullet Z^c + b_5X \bullet Z \qquad \textbf{Equation 6}$$

ΔR^2 between Equations 5 and 6 constitutes a test of moderation for investigators facing high multicolinearity (r_{xz}) and possible nonlinear relationships between Y and X or Y and Z.

Error 4: Causal Sequencing

This error occurs when the causal order is incorrectly specified, i.e., confusing X→Y with Y→X. While tests of simple linear relationships between X and Y are not affected by which is designated the cause and which the effect, this is not true with MMR. Reexamination of moderation's conceptual definitions above reveals that some do not specify an X→Y or Y→X causal order. Others clearly specify a predictor "X" and criterion "Y." Regardless, tests of whether Z moderates X→Y and Y→X differ both phenomenologically and methodologically.

Landis and Dunlap (2000) demonstrated F statistics calculated to test H_0: $\Delta R^2 = R^2_{mult} - R^2_{add}$

= 0 are not equal for ΔR^2 for the following MMR analyses:

$$\hat{Y} = b_0 + b_1X + b_2Z;\ R^2_{add} \qquad \textbf{Equation 7}$$

$$\hat{Y} = b_0 + b_1X + b_2Z + b_3XZ;\ R^2_{mult} \qquad \textbf{Equation 8}$$

and

$$\hat{X} = b_0 + b_1Y + b_2Z;\ R^2_{add} \qquad \textbf{Equation 9}$$

$$\hat{X} = b_0 + b_1Y + b_2Z + b_3YZ;\ R^2_{mult} \qquad \textbf{Equation 10}$$

Landis and Dunlap demonstrated MMR may yield different results because investigators who choose the incorrect X→Y causal sequence will not test the interaction term associated with the true underlying interaction phenomena (i.e., X•Z versus Y•Z). Extending Harris' (1997) labels, we would consider this a Type IV error, where incorrect conceptualization leads to a test of the wrong question.

This error was difficult to detect in the articles reviewed as it relies on deep knowledge of the target phenomena for each article to determine whether reverse causal ordering is a reasonable alternative. Many MIS authors explicitly recognized the emergent nature of IT phenomena. For example, Harrison et al. provided a feedback loop recognizing that not only do attitudes, subjective norms, and perceived control impact adoption intentions, but adoption and actual control of an innovation impacts the attitudes, subjective norms, and perceived control impacting future adoption intentions.

In contrast, Armstrong and Sambamurthy (1999) modeled recursive main effects in a model of relationships between senior leadership knowledge, systems of knowing, and IT assimilation. Strategic vision was examined as a moderator, although relationships were tested in only one direction. Articles illustrating best practices for this error all established a single X→Y causal order on the basis of experimental manipulation of X (Ahituv et al 1998; Keil at al 2000) or use of longitudinal designs (where future observations of Y could not have caused past observations of X; e.g., Devaraj and Kohli 2000).

Solution. Investigators need to be aware of theoretical rationale justifying the X→Y or Y→X causal orders. The most severe consequence of failure to thoroughly explore justifications for alternate causal orders, i.e., misaligning research design with the true latent causal sequence, would result in a literature littered with evidence supporting an X•Z (or Y•Z) interaction effect when in fact that interaction cannot exist because Y→X (or X→Y).

Absent strong theoretical rationale, examining both possible moderator effects (X•Z and Y•Z) in the context of relationships with other predictor variables seems to be the best course of action available. Simultaneously, MIS investigators should perform exploratory analyses aimed at developing strong theoretical rationale to guide future analyses (Glaser and Strauss 1967).

Error 5: Low Power of Random Effects Designs

Recall random effects designs occur when variation in treatment levels or values of the independent variable are assumed to be randomly distributed in the population of interest. Investigators using a fixed effect design control who is exposed to what levels of treatments on the independent variable and generally do so in a way that maximizes statistical power (i.e., the investigator is conducting a controlled experiment). The former occur most frequently in survey research where investigators measure independent variables using survey instruments. Assumptions that either (1) X and Z are normally distributed or (2) residual prediction error *e* is normally distributed are necessary but not sufficient conditions for common parametric tests of statistical significance (e.g., H_0: $\Delta R^2 = R^2_{mult} - R^2_{add} = 0$).

Schepanski (1983) considered three investigators examining whether an X-Y relationship is moderated by Z. The first found X and Z take on the values 1, 2, 3, 4, 5, 6, 7, 8, and 9 and a sample of N = 81 observations was obtained for every possible X,Z combination. If the true latent causal process is Y = X•Z and all variables are measured without error, the MMR effect size is $\Delta R^2 = R^2_{mult} - R^2_{add} = .061$. Alternatively, the second investigator for some bizarre reason obtains a sample of N = 81 paired X,Z observations with values (1,9), (2,8), (3,7), (4,6), (5,5), (6,4), (7,3), (8,2), and (9,1) occurring with equal frequency (i.e., the off-diagonal cells in a 9 × 9 experimental design). In this instance, $R^2_{add} = 0$ and $\Delta R^2 = 1.0$. Finally, a third investigator obtains a sample containing 81 observations drawn from cells in which Z + X = 9, 10, or 11 (i.e., the off-diagonal cells in a 9 × 9 experimental design and immediately adjacent cells). In this instance, $\Delta R^2 = .39$. These results demonstrate X,Z combination frequencies directly influence the sample size needed to reject H_0: $\Delta R^2 = 0$ when moderation is present.

Schepanski explained these outcomes in terms of the power of additive models when data exhibit conditionally monotone independent → dependent variable relationships. An additive model will perfectly explain data in which all observations exhibit strict dominance, i.e., where one member of every pair of observations "possess higher values on one or more independent variables and equal values" on all other independent variables (Schepanski 1983, p. 505). The three hypothetical investigators described above obtained different ΔR^2 effect sizes because the data sets differed in proportion of paired data points exhibiting strict dominance. At one extreme, the second investigator's data set contained no strictly dominant pairs of observations; none of the observations exhibited strict dominance relative to any other observation, and the additive model exhibited no predictive power ($R^2_{add} = 0.00$). At the other extreme, 18 percent of the paired observations exhibited strict dominance in the third investigator's data set.

Many articles in our sample reported very small sample sizes; however, Hardgrave et al.'s (1999) exploration of prototyping strategy seems particularly compelling. They surveyed 133 firms in a random effects field design about 168 different prototyping projects and used moderated regression analysis to evaluate 15 hypothesized moderator effects. The number of observations available for analysis varied from 91 to 111 (presumably due to missing data) and the inclusion of 16 main effects (15 hypothesized moderators and type of prototype employed) and one interaction effect consumed 17 degrees of freedom. Hence, F-test of H_0: $\Delta R^2 = 0$ for the interaction effects yielded a df range from 1,74 to 1,94. The average ΔR^2 reported was .0261. Given the largest R^2_{mult} reported by Hardgrave et al. was $R^2_{mult} = .131$ and obtaining the critical value at $\alpha = .05$ of $F_{1,80} = 3.84$ (i.e., conservatively using the largest df = 111 – 17 = 94 reported), solving the formula

$$F_{1,80} = \frac{\Delta R^2}{(1 - R^2_{mult}) / (80)}$$

for ΔR^2 indicates these authors' analyses at best would only have rejected H_0: $\Delta R^2 = 0$ when observed $\Delta R^2 \geq .0363$, which is more than50 percent larger than the average ΔR^2.

The question remaining is, what ΔR^2 should Hardgrave et al. have expected if true moderation effects were occurring? If the expected $\Delta R^2 <$ .0363, then failure to reject H_0: $\Delta R^2 = 0$ would be expected as the sample size and observed R^2_{mult} only permitted detection of moderator effects which yield $\Delta R^2 \geq .0363$.

Solution. At least two implications can be drawn for MIS investigators. First, before initiating a study in which moderation is hypothesized, investigators should estimate the frequency with which X and Z assume different values and forecast the expected ΔR^2 effect size.[5] Solving Equation 3 for N will find the minimum sample size needed to detect any true interaction effect.

[5]Note any estimate of effect size will have to take into account reliability of X,Z and the XZ product term using Busemeyer and Jones' (1983) correction, which is described in the section discussing Error 8.

As part of a program of research examining prototyping, Hardgrave et al. should count the relative proportion of strictly dominant paired observations in their data. Given most phenomena were measured using seven-point Likert scales, Hardgrave and his colleagues would then generate paired observations with this particular proportion of strictly dominant pairs in a Monte Carlo simulation of a 3 × 7 design (three types of prototype strategy by seven possible levels of each moderator). The ΔR^2 obtained from this simulation (corrected for measurement error as described in the solution to Error 8) would then be plugged into Equation 3 along with the average R^2_{mult} reported in the original Hardgrave et al. effort and the critical value of F to determine the minimum sample size needed to detect this expected ΔR^2 effect size.

Second, authors should choose X,Z combinations that maximize statistical power (P{reject H_0} when H_0 is false for tests of H_0: $\Delta R^2 = 0$). McClelland and Judd (1993) demonstrated data sets containing observations drawn only from X,Z combinations of (1,1), (1,9), (9,1), and (9,9) in Schepanski's example maximized statistical power. This suggests investigators must take special care in trimming any outlier observations from the data. Given McClelland and Judd's demonstrated observations drawn from extreme X,Z combinations maximize ΔR^2, investigators who incorrectly label outlier observations as having been drawn from some population other than the population of interest are effectively decreasing expected ΔR^2 effect size and increasing the sample size required to reject H_0: $\Delta R^2 = 0$.

Finally, careful readers will note a subtle distinction in our discussion of Error 5, i.e., the distinction between model testing versus maximizing Y prediction accuracy (Birnbaum 1973, 1974). The percent of strictly dominant paired cells in a study's design will determine both the incremental increase in prediction accuracy by the multiplicative model and whether the additive model is rejected (Aguinis 1995). As noted by Schepanski, when the true latent model is multiplicative, knowledge of that fact will add minimally to prediction accuracy in a population containing mostly strictly dominant pairs of observations (e.g., an additional 6.1 percent of the variance for the first investigator above). However, in those cases in which the additive and multiplicative models yield different $\hat{Y}_i$ estimates, $\hat{Y}_{add_i}$ will be very different from $\hat{Y}_{mult_i}$ and prediction error for the additive model ($\hat{Y}_{add_i} - Y_i$) will be much larger than for the multiplicative model ($\hat{Y}_{multi} - Y_i$). Hence, while incremental variance explained may be minimal in some populations of X, Z, and Y observations, the investigator (the first investigator in the examples above) risks making a small number of very severe prediction errors when embracing an incorrect additive model simply because it is more parsimonious and explains the vast majority of Y variance. Investigators aligning their research designs with the phenomenon of interest must weigh both the relative frequency and severity of errors before endorsing the simpler additive model.

Measurement and Scaling Issues

Errors 6, 7, 8, and 9 occur due to issues involving scale coarseness, nonlinear transformations, measurement error, and use of different subgroup measurement models.

Error 6: Dependent Variable Scale Is Too Coarse

When X and Z take on multiple possible values, a true model Y = X•Z will yield a latent dependent outcome Y that often contains more possible levels than investigators used in measuring Y. For example, if X and Z are phenomena measured on five-point interval scales, Y could have at least seven different values (e.g., if X and Z range from –2 to +2, Y = X•Z takes on the values of –4, –2, –1, 0, 1, 2, 4) and at most 25 different values. Subjects faced with reporting Y responses on a five-point Likert scale must somehow reduce their latent 7- to 25-point dependent Y response into the relatively coarse five-point overt response format.

Russell and Bobko (1992) found subjects in this exact scenario using the model Y = X•Z and facing a 150-point overt response scale yielded a ΔR^2 MMR effect size that was 97 percent larger than subjects faced with placing overt Y responses on a traditional five-point Likert scale. It is important to place this result in context of Likert's (1932) oft replicated finding that increasing the number of response categories beyond five to seven does not yield substantial gains in observed reliability (cf. Cicchetti et al. 1985). While reliability may not change, construct validity of the dependent Y measure may. Russell and Bobko's findings suggest investigators using Likert scales that are too coarse relative to latent Y construct domains will dilute construct validity of their Y operationalizations and attenuate ability to detect true moderation relationships. MIS investigators who do not determine the number of meaningfully different levels of Y occurring from Y = X•Z run the risk of severely attenuating observed ΔR^2.

In all, 20 articles in our sample used dependent measures that were too coarsely scaled (ranging from five- to ten-point Likert scales). In one potential example of this error, Agarwal and Prasad (1999) reported tests of three interactions that were theoretically related to intentions to use IT innovations: perceived usefulness × personal innovation, perceived ease of use × personal innovation, and compatibility × personal innovation. Measures obtained on an individual difference characteristic, three perceptual variables, and the dependent variable all used seven-point Likert item response scales. Hence, subjects were potentially faced with portraying a latent 7 × 7 = 49 level latent dependent response on a seven-point scale used to measure intention to use an IT innovation. Agarwal and Prasad's MMR analysis found only a compatibility × personal innovation effect statistically significant. Russell and Bobko's findings suggest a 49-point scale to measure intention to use IT innovations could have caused Agarwal and Prasad to enjoy at least a 97 percent increase in effect size for the perceived usefulness × personal innovation and perceived ease of use × personal innovation effects if these effects were actually present.

Illustrating a potential best practice, Keil et al. investigated the moderating effect of national culture on relationships between risk propensity, level of sunk cost, and risk perception. In their study, risk perception was operationalized using a 100-point scale.

Solution. The solution requires investigators to identify *a priori* the expected number of distinct X and Z values (i.e., $\#_x$, $\#_Y$) that might occur and select a Y measurement scale portraying all $\#_x \bullet \#_Y$ possible values. Arnold (1981) reported pilot efforts that might be used to establish $\#_x$ and $\#_Y$ in the first field test to unambiguously support Vroom's (1964) original multiplicative expectancy theory formulation. Cautious investigator will operationalize Y as a continuous variable (i.e., one that can take on an infinite number of values).

Error 7: Nonlinear Monotonic Transformations on Y, X, and Z

A number of assumptions must be met to use Equation 3 to test H_0: $\Delta R^2 = 0$. In a random effects design, one must assume X and Z are distributed multivariate normal or that prediction error (*e*) is normal with a constant standard deviation across all predicted Y values (commonly referred to as homoskedasticity). A number of transformations are available to convert observations in such a way that they less severely violate one or more of these assumptions. For example, statistical texts routinely reference log transformations to make a positively skewed distribution appear more bell shaped or normal (Winer 1974). Other common nonlinear transformations include use of arc-sine transformations on percentage data, square roots, and Fischer's z transformation on Pearson product moment correlations.[6]

Theoretical rationale exists for nonlinear interval scale transformations in a number of arenas (e.g., Stevens 1958). We are unaware of any theories or models in applied management research that

[6]See Bartlett (1947) for a discussion of log, arc-sine, and square root transformations.

provide strong theoretical rationale for nonlinear interval scale transformations.[7] Statistical elegance (i.e., meeting parametric assumptions) appears to be the major purpose of these transformations. Unfortunately, severe unintended consequences can occur.

Specifically, Busemeyer and Jones (1983) demonstrated "when it is theoretically permissible to monotonically transform the criterion variable, then hierarchical regression analysis cannot yield an interpretable test of the multiplicative versus additive structural model" (p. 555). They provided examples showing how data derived from a truly additive model (e.g., $\hat{Y} = b_0 + b_1X + b_2Z$) can be monotonically transformed in such a way that H_0: $\Delta R^2 = 0$ will be rejected and how data derived from a truly multiplicative model can be monotonically transformed in such a way that H_0: $\Delta R^2 = 0$ will not be rejected. MMR results do not provide a reliable index of moderation effects when Y has been subjected to monotonic transformation. Birnbaum (1973, 1974) demonstrated the same problems occur when X and Z are subjected to nonlinear, monotonic transformations.

Only one article in our sample reported a nonlinear transformation. Harrison et al. investigated the moderating effect of organizational size on the relationship between attitudes, subjective norms, perceived control, and decisions to adopt. However, they performed a logarithmic transformation of their organizational size variable. Analysis produced a significant ΔR^2 and they concluded organizational size does moderate the relationship between their independent and dependent variables. Unfortunately, as noted by the Busemeyer and Jones quote above, Harrison et al.'s significant ΔR^2 is not interpretable: no conclusion can be drawn from their analyses about organizational size and moderation.

Solution. Russell and Dean (2000) recently applied bootstrapping procedures to estimate confidence intervals around ΔR^2 without transforming the dependent variable or making parametric assumptions. Using examples involving positively skewed dependent variables drawn from compensation research, Russell and Dean found the preferred monotonic transformation (i.e., a log transformation) severely decreased estimates of true moderator effects using moderated regression procedures in a Monte Carlo simulation. MMR ΔR^2 moderator effect sizes were substantially better estimates of the true latent moderator effect (i.e., larger by a multiple of 2.6 to 534) when estimated using a simple percentile bootstrap procedure in the original, untransformed (positively skewed) data.[8]

Conclusions regarding the presence or absence of a true moderator effect using simple bootstrap procedures were unaffected by violations of parametric assumptions in the original, positively skewed data. Conclusions when moderated regression analysis was performed on a log Y severely increased frequency of Type II errors. Hence, Harrison et al. could have arrived at an interpretable test of H_0: $\Delta R^2 = 0$ if they had followed this bootstrap procedure. It remains to be seen whether bootstrap procedures for estimating ΔR^2 confidence intervals exhibit the same power in circumstances where characteristics of the Y distribution suggest a monotonic transformation other than a log Y. Regardless, applied behavioral science investigators should never use MMR when Y has been subjected to monotonic transformations absent some strong theoretical (i.e., not statistical) justification.

Error 8: Influence of Measurement Error on X•Z

Well-trained MIS investigators conducting programmatic research usually estimate the sample size necessary to detect the effect of interest (i.e., reject H_0 at $p < .05$). Using a 5 × 5 experimental design to gather Y observations from subjects who were known to generate them from a Y = X•Z

[7]The only exception we are familiar with is the notion of marginal decreasing utility of money from labor economics.

[8]Note confidence intervals for parameters estimated using PLS are estimated using bootstrap procedures.

model, Russell and Bobko found the average $\Delta R^2 = .03$ when a five-point Likert scale was used to measure Y. Assuming $\Delta R^2 = .03$, a sample of N = 96 would have been needed to reject H_0: $\Delta R^2 = 0$, where N = 96 is derived by solving Equation 3 (reprinted below) for N:

$$F_{1,N-3} = \frac{\Delta R^2}{(1 - R^2_{mult}) / (N - 3)}$$

where $\Delta R^2 = .03$,
$R^2_{mult} = .25$, and
$F_{1,N-3} = 3.84$

The ΔR^2 used in this example already reflects measurement error, i.e., $\Delta R^2 = .030$ was Russell and Bobko's observed ΔR^2 derived using Y measures that contained measurement error. MIS investigators examining moderation phenomena for which estimates of ΔR^2 have not been reported in the literature will have to estimate ΔR^2 by simulating X and Z distributions, using them to create Y = X•Z, and finally deriving ΔR^2 and the attendant N needed to detect it. However, ΔR^2 obtained from simulation data must be attenuated for measurement error in order to accurately approximate $E(\Delta R^2)$, and hence the estimate of N needed to reject H_0: $\Delta R^2 = 0$.

Most of the reliabilities reported in our sample fall above Nunnally's (1967) $\alpha \geq .70$ rule of thumb, although seven do not. For example, McKeen et al. (1994) examined the relationship between user participation and user satisfaction, task complexity, system complexity, user influence, and user-developer communication as moderators. Reliabilities for system complexity ($\alpha = .65$) and user-developer communications ($\alpha = .54$) fell below the .70. Further, the authors rejected H_0: $\Delta R^2 = 0$ for two (task complexity and system complexity). It is possible the two insignificant findings were due to a reduction in observed ΔR^2 due to measurement error.

Boudreau et al. (2001) examined MIS research from 1997 through 1999 and concluded at least 20 percent (depending on the journal) of published empirical work failed to report reliability measures. Articles in our sample using perceptual measures all reported reliabilities (in varying degrees of detail): the average sample size for those studies was $\bar{N} = 255.2$ and the weighted average reliability (weighted by sample size) across all reliabilities reported was $\bar{\alpha} = .824$. Authors and reviewers should insure reliabilities are reported to assist future assessments of measurement error's impact on required sample sizes.

Solution. Busemeyer and Jones also developed a method of correcting expected MMR effect size for measurement error in X and Z. $\rho_{X \bullet Z} = 1.00$ if MIS investigators use fixed effects designs in which there is no measurement error in X or Z. Alternatively, the MIS investigator using a random effects design and questionnaire measures will likely have operationalizations of X and Z (i.e., X and Z scale scores) containing measurement error. The MIS investigator can simulate X and Z observations to (1) estimate expected interaction effect size (ΔR^2) in the absence of measurement error (described in solutions to Errors 5 and 7), (2) plug X and Z reliability estimates obtained from the literature into Equation 11, (3) plug that result into Equation 12 to estimate the expected ΔR^2 obtained under actual research conditions in which measurement error is present.

$$\rho_{x \bullet z} = \frac{(\rho_x \bullet \rho_z)}{1 + \rho^2_{x,z}} \qquad \textbf{Equation 11}$$

$$\Delta\rho^2 = \frac{\rho_{x \bullet z}[b_3^2 \bullet s^2_{xz}]}{s^2_y} \qquad \textbf{Equation 12}$$

Where $\rho_{X \bullet Z}$ = reliability of the X•Z product term
ρ_X = reliability of X
ρ_Z = reliability of Z
$\rho_{X,Z}$ = simple correlation between X and Z
b_3 = regression coefficient for the product term in Equation 2
s^2_{xz} = variance of the X•Z product term
s^2_y = variance of the dependent variable Y

Using this expected ΔR^2 estimate in Equation 3 will yield a more accurate estimate of the sample size needed to reject H_0: $\Delta R^2 = 0$ if in fact $Y = X \bullet Z$.

Finally, Chin et al. (1996) demonstrated how PLS derives estimates of regression coefficients after correcting for X and Z internal consistency reliability estimates. We would expect PLS results to converge with MMR results that have been corrected for unreliability using Busemeyer and Jones' formula if initial estimates of X and Z reliabilities are the same.

Error 9: Gamma Differences in PLS

A final measurement issue is unique to the use of PLS. The PLS technique differs from MMR in that it provides for concurrent estimation of the structural and measurement models. In doing so, it derives factor scores (by summing the products of PCA factor analysis loadings and subjects' item responses) as best estimates of latent constructs.

Six studies in our sample used PLS to test moderation hypotheses in which the moderator Z was a dummy coded variable capturing membership in one of two or more groups. Further, three out of the five moderation studies published in *MIS Quarterly* and *Information Systems Research* in 2000 used this method (e.g., Keil et al. 2000; Venkatesh 2000; Venkatesh and Morris 2000).

Tests for moderation using PLS require separating samples into groups where membership is based on some level of the hypothesized moderator variable. Separate analyses are run for each group and path coefficients are generated for each subsample. Path coefficients are then compared to determine whether the relationship between some set of predictors X and criteria Y depended on subgroup membership Z. In a recent example, Keil et al. derived separate PLS estimates for latent structural relationships between risk propensity, risk perceptions, sunk costs, and project escalation for three samples drawn from different cultures. Comparing path coefficients across subsamples indicated culture moderated the relationship between risk propensity and risk perception.

The comparison of the same path coefficient in two subsamples (Chow 1960) is computationally the same as rejecting H_0: $\Delta R^2 = 0$ in an MMR analysis in which X is some continuous predictor and Z is a dummy coded nominal variable (Bobko 1995, pp. 228-229). Problems occur when PLS derives new factor loadings and weights in separate analyses conducted in each subsample. The construct-level scores are subsequently estimated using different item weights in each subsample. For example, Kiel et al. compared path coefficients in models in Singapore, Finland, and the Netherlands. Risk perception was a composite of four questions. Kiel et al. did not report item weights, although their Table 3 shows that the factor loadings for the risk perception items were different in each subsample. At the extreme, item 2 loadings varied from .57 to .90. Loading variability suggests PLS also varied item weights, causing estimates of the risk perception construct to be created from different weighted combinations of the four items in each subsample and influencing statistical tests for differences in path coefficients. Simply stated, risk perception scores derived in this manner have substantially different meanings for observations drawn from Finland, the Netherlands, and Singapore. Path coefficients may differ significantly across countries when risk perception is constructed from a different weighted sum of the four items and not differ significantly when risk perception is a simple sum of the four item responses (or vice versa).

This is one of many examples in MIS research using PLS to examine differences in path coefficients across groups. In these instances, PLS confounds true differences in path coefficients with differences in latent construct composition (i.e., different factor loadings), preventing any interpretations of PLS results bearing on the hypothesized moderation effect. Interested readers should see discussions by Rice and Contractor (1990), Schmitt (1982), and Schmitt et al. (1984) of gamma differences in latent factor structure between two administrations of the same instrument.[9]

---

[9]Alpha change occurs when some true change has occurred between administrations of some measure. Beta change occurs when no true change occurred, although a difference in observed scores occurs due to a change in scaling (i.e., commitment previously viewed as a "3" on a Likert scale is now viewed as "3.5").

Table 3. Factor Loadings for Risk Perception (Keil et al. 2000)				
	Full Sample (n = 536)	**Finland (n = 185)**	**Netherlands (n = 121)**	**Singapore (n = 230)**
Item1	.88	.91	.75	.88
Item2	.86	.90	.57	.88
Item3	.71	.57	.79	.71
Item4	.69	.72	.76	.69

Solution. Two possible solutions exist. First, if the two groups reflected in the dummy coded Z variable are independent, investigators should test the null hypothesis that inter-item covariance matrices within scales are equal using Box' M test of equal covariance matrices. Duxbury and Higgins (1991) performed a variation of this, inferring measurement equivalence based on the absence of mean differences (using unpaired t tests) between men and women. Box' M test of equal covariance matrices between scales scores found significant differences, although Duxbury and Higgins appropriately interpreted this as being due to male-female differences in associations between constructs. Unfortunately, the presence or absence of differences in scale score means as determined by the unpaired t tests is irrelevant to the construct validity issue: males and females might or might not exhibit true differences on Duxbury and Higgins' constructs. The real issue is whether the construct contents as determined by item loadings within scales are the same. Unfortunately, Duxbury and Higgins' did not compare covariance matrices at the item level, which would have determined the degree to which scale scores reflected similar latent constructs for males and females. Note, comparisons of covariance matrices across all items could reject the null hypothesis of equal covariance matrices due to differences in construct content or differences in relationships among constructs (i.e., the measurement model and structural model).

If the two groups are not independent (e.g., two administrations of a single measure to the same sample at different points in time), investigators' should derive the coefficient of concordance described by Harman (1976) to assess similarity of factor loadings. Similarity in item correlation matrices or factor loadings will permit investigators to assess whether latent constructs being measured in the two groups are the same. In this instance and only this instance can the investigator then derive scale scores in the same manner for observations in both groups.

If there is no evidence suggesting similarity in the latent construct domain across the Z groups, PLS (and MMR) could still be performed, although traditional moderator interpretations cannot be drawn. Moderation may be present, though the *it* in "it all depends on…" is fundamentally different. Instead of *it* referring to how the X→Y relationship varies across groups Z = 0 and 1, *it* refers to the fact that what constitutes "X" fundamentally differs across the two groups (i.e., observed "X" in group 1 may tap latent construct X, although observed "X" in group 2 taps latent construct Q). Observed "X"→Y relationships may vary for Z = 0 versus 1, although differences in these relationships really mean the X→Y relationship in group 1 differs from the Q→Y relationship in group 2.

Conclusions

Tests for moderation are a significant part of the growing body of empirical research findings in MIS. While many MIS investigators are aware of a number of the issues presented here, mixed results containing substantial numbers of Type I and Type II errors will occur less frequently if authors, reviewers, and editors are more aware of

1 **Table 4. Guidelines for Authors and Evaluators**

2	**Error**	**Advice to Authors**	**Advice to Evaluators**	**Result**
3		Authors should take care to describe the type of moderation they are hypothesizing. Specifically authors need to be certain whether it is the strength or the nature of the $X \rightarrow Y$ relationship that depends on the moderator variable (Z) and then match the analysis method to this conceptual definition.	Where authors are interested in differences in the strength of the $X \rightarrow Z$ relationship depending on levels of Z, the issues addressed in this manuscript are all relevant.	Matching analysis method to the correct hypothesized interactions avoids Type IV errors where incorrect conceptualization leads to a test of the wrong question.
4	1	Report effect size in the form of ΔR^2 or an equivalent measure (such as η^2).	Without ΔR^2 (or an equivalent measure) no conclusions can be drawn about effect size.	This is important in helping readers understand the contribution of the study in hand but also in helping MIS researchers be more aware of the overall role moderation plays in understanding MIS issues.
5	2	Interpret main effects only when moderating effects are insignificant.	No conclusions can be drawn about main effects in the presence of moderating effects.	Both Type I and Type II errors can be avoided.
6	3	Report correlation matrix. Report application of equations 4-6 to partial out any X^2 effects when X and Z are highly correlated.	Failure to partial out X^2 effects could cause researchers to conclude a moderation effect exists when in fact it is a nonlinear relationship between X and Y in disguise. This is especially a concern when X and Z are highly correlated.	Type I errors can be avoided.
7	4	Report evidence to clearly establish causal ordering or results from investigating both $X \rightarrow Y$ and $Y \rightarrow X$.	Authors who fail to clearly establish causal order may be testing the wrong question. The ordering can be established theoretically or by research design.	Clearly establishing causal order (or examining effects in both causal sequences) avoids Type IV errors, the potential error of testing the wrong questions.
8	5	Report power analysis and needed sample size.	In the case of insignificant findings, evaluate whether or not the sample size is sufficient to find moderating effects when they are present.	Type II errors can be avoided.

Table 4. Guidelines for Authors and Evaluators (Continued)

Error	Advice to Authors	Advice to Evaluators	Result
6	Report all scales.	The scale of the dependent measure should reflect the product of the independent and moderating variables.	Type II errors can be avoided.
7	Report all transformations, nature of transformation, and rationale.	Nonlinear transformations of predictor, criterion, or moderator variables make the comparison of multiplicative and additive models uninterpretable.	Both Type I and Type II errors can be avoided.
8	Report scale reliabilities.	Low reliabilities can attenuate ΔR^2.	Type II errors can be avoided.
9	Report item weights when using PLS. Also, report Box' M and/or coefficient of concordance.	Subgroups cannot be compared without evidence that they do not vary significantly in construct score weighting.	Type I errors can be avoided.

the nine errors and solutions described above. Importantly, several of these errors can be avoided only if authors follow and editors enforce certain reporting standards. We provide reporting guidelines and advice for evaluators in Table 4. This table also summarizes the consequences of these errors (either in the form of erroneously rejecting the null hypothesis, erroneously accepting the null hypothesis, or failing to test the correct question). This is important to note because there is a subtle but important distinction. Errors 2, 3, 7, and 9 can result in Type I error (i.e., false positive results) and consequently results derived from studies having committed these errors are potentially invalid. In contrast, studies committing errors 5, 6, or 8 may be committing Type II errors (i.e., false negative results) when moderation is present. Error 4 can result in the wrong question being investigated, leading to Type I or Type II errors.

For researchers beginning a new study, the message is clear. Errors 2, 3, 4, 7, and 9 must be avoided. Further, errors 5, 6, and 8 should be avoided. If they cannot be avoided (for example, the researcher has calculated the required sample size and it is unattainable), then the researcher should be aware the effort is risky: the likelihood of detecting the true moderation effect is very low. Error 1 may result in Type I or Type II errors, and its reporting directly effects our ability to accumulate findings. This error can and should always be avoided.

While some of these errors have been made for decades in applied behavioral science research, the most recent manifestation occurred with the advent and increasing popularity of PLS applications in MIS research. It was not our intention to imply PLS analysis is inappropriate. Use of PLS when fundamental differences in a latent construct content exist between groups can lead to severe misinterpretations regarding the presence or form of any moderator relationships.

In sum, researchers can lower the cost of and increase the speed with which new MIS knowledge is generated by avoiding the problems described above. MIS researchers are forced to make decisions balancing study generalizability against the control exercised over research environments, i.e., to balance the relevance of studies against the rigor with which they are conducted. Studies involving tests of moderation will be more powerful and rigorous when the nine errors reviewed above are minimized. Investigator decisions about which statistics to report, how to interpret them, designs and analysis techniques to apply, and how to operationalize constructs of interest in the search for moderation effects directly influence the accuracy of subsequent results and conclusions drawn.

Acknowledgements

Comments by Laku Chidambaram, Wynne Chin, Carol Saunders, Jane Webster, and Bob Zmud on an earlier version of this paper are greatly appreciated. Order of authorship is alphabetical and reflects the authors' equal contributions to the study.

References

Agarwal, R., and Prasad, J. "A Conceptual and Operational Definition of Personal Innovativeness in the Domain of Information Ttechnology," *Information Systems Research* (9:2), 1998, pp. 204-215.

Aguinis, H. "Statistical Power Problems in Multiple Moderated Regression in Management Research," *Journal of Management* (21:6), 1995, pp. 1141-1158.

Ahituv, N., Igbaria, M., and Sella, A. "The Effects of Time Pressure and Completes of Information on Decision Making," *Journal of Management Information Systems* (15:2), 1998, pp. 153-172.

Aiken, L. S., and West, S. G. *Multiple Regression: Testing and Interpreting Interactions*, Sage Publications, Newberry Park, CA, 1996.

Armstrong, C. P., and Sambamurthy, V. "Information Technology Assimilation in Firms: The Influence of Senior Leadership and IT Infrastructures," *Information Systems Research* (10:4), 1999, pp. 304-327.

Arnold, H. J. "Moderator Variables: A Clarification of Conceptual, Analytic, and Psychometric Issues," *Organizational Behavior and Human Performance* (29:2), 1982, pp. 143-174.

Arnold, H. J. "Test Moderator Variable Hypothesis: A Reply to Stone and Hollenbeck," *Organizational Behavior and Human Performance* (34), 1984, pp. 214-224.

Arnold, H. J. "A Test of the Multiplicative Hypothesis of Expectancy Valence Theories of Work Motivation," *Academy of Management Journal*, (24) 1981, pp. 128-141.

Banerjee, D., Cronan, T. P., Jones, T. W. "Modeling IT Ethics: A Study of Situational Ethics," *MIS Quarterly* (22:1), 1998, pp. 31-60.

Banker, R. D., and Slaughter, S. A. "The Moderating Effects of Structure on Volatility and Complexity in Software Enhancement," *Information Systems Research* (11:3), 2000, pp. 219-240.

Baron, R. M., and Kenny, D. A. "The Moderator-Mediator Variable Distinction in Social Psychology Research: Conceptual, Strategic, and Statistical Considerations," *Journal of Personality and Social Psychology* (51:6), 1986, pp. 1173-1182.

Baroudi, J. J., and Orlikowski, W. J. "The Problem of Statistical Power in MIS Research," *MIS Quarterly* (13:1), 1989, pp. 87-106.

Bartlett, M. S. "The Use of Transformations," *Biometrics* (3), 1947, pp. 22-38.

Birnbaum, M. H. "The Devil Rides Again: Correlation as an Index of Fit," *Psychological Bulletin* (79), 1973, pp. 239-242.

Birnbaum, M. H. *Measurement, Judgement, and Decision Making*, Academic Press, San Diego, CA, 1998.

Birnbaum, M. H. "Relationships Among Models of Salary Bbias," *American Psychologist* (40), 1985, pp. 862-866.

Birnbaum, M. H. "Reply to the Devil's Advocates: Don't Confound Model Testing with Measurement," *Psychological Bulletin* (81), 1974, pp. 854-859.

Birnbaum, M. H. "To Resolve Fechner vs. Stevens: Settle the Dispute Concerning 'Ratios' and 'Differences,'" *Behavioral and Brain Sciences* (12), 1989, pp. 270-271.

Bobko, P. *Correlation and Regression: Principles and Applications of Industrial/Organizational Psychology and Management*. McGraw-Hill Inc., New York, 1995.

Bobko, P. "A Solution to Some Dilemmas When Testing Hypotheses About Ordinal Interactions," *Journal of Applied Psychology* (71), 1986, pp. 323-326.

Boudreau, M. C., Gefen, D., and Straub, D. W. "Validation in Information Systems Research: A State-of-the-Art Assessment," *MIS Quarterly* (25:1), 2001, pp. 1-16.

Busemeyer, J. R., and Jones, J. E. "Analysis of Multiplicative Combination Rules When the Causal Variables Are Measured with Error," *Psychological Bulletin* (93), 1983, pp. 549-562.

Chin, W. W., Marcolin, B. L., and Newsted, P. R. "A Partial Least Squares Latent Variable Modeling Approach for Measuring Interaction Effects: Results from a Monet Carlo Simulation Study and Voice Mail Emotion/Adoption Study," in *Proceedings of the Seventeenth International Conference on Information Systems*, J. I. DeGross, S. Jarvenpaa, and A. Srinivasan (eds.), Cleveland, OH, 1996, pp. 21-41.

Choe, J. M. "The Relationships Among Performance of Accounting Information Systems, Influence Factors, and Evolution Level of Information Systems" *Journal of Management Information Systems* (12:4), 1996, pp. 215-239.

Chow, G. "Test of Equality between Sets of Coefficients in Two Linear Regressions," *Econometrica* (28:3), 1960, pp. 591-604.

Cicchetti, D. V., Showalter, D., and Tyrer, P. J. "The Effect of Number of Rating Scale Categories on Levels of Interrater Reliability: A Monte Carlo Investigation," *Applied Psychological Measurement* (9), 1985, pp. 31-36.

Cohen, J. "Partial Products Are Interactions µ Partial Powers Are Curve Components," *Psychological Bulletin* (85), 1978, pp. 858-866.

Cohen, J., and Cohen, P. *Applied Multiple Regression/Correlation Analysis for the Behavioral Sciences* (2nd ed.), Lawrence Erlbaum, Hillsdale, NJ, 1983.

Cortina, J. M. "Interaction, Nonlinearity, and Multicolinearity: Implications for Multiple Regression," *Journal of Management* (19), 1993, pp. 915-922.

Devaraj, S., and Kohli, R. "Information Technology Payoff in the Health-Care Industry: A Longitudinal Study," *Journal of Management Information Systems* (16:4), 2000, pp. 41-67.

Duxbury, L. E., and Higgins, C. A. "Gender Differences in Work-Family Conflict," *Journal of Applied Psychology* (76:1), 1991, pp. 60-74.

Duxbury, L. E., Higgins, C. A., and Mills, S. "After-Hours Telecommuting and Work Family Conflict: A Comparative Analysis," *Information Systems Research* (3:2), 1992, pp 173-190.

Fritz, M. B. W., Yarasimhan, S., and Rhee, H.-S. "Communications and Coordinations on the Virtual Office" *Journal of Management Information Systems* (14:4), 1998, pp. 7-28.

Gefen, D., Straub, D. W., and Boudreau, M. C. "Structural Equation Modeling and Regression: Guidelines for Research Practice," *Communications of the AIS* (1:7), 2000, pp. 1-78.

Glaser, B. G., and Strauss, A. L. *The Discovery of Grounded Theory: Strategies for Qualitative Research*, Aldine Publishing, New York, 1967.

Grover, V., Cheon, M. J., and Teng, J. T. C. "The Effect of Service Quality and Partnership on the Outsourcing of Information Systems Functions," *Journal of Management Information Systems* (12:4), 1996, pp. 89-116.

Hardgrave, B. C., Wilson, R. L., and Eastman, K. "Toward a Contingency Model for Selecting an Information System Prototyping Strategy," *Journal of Management Information Systems* (16:2), 1999, pp. 113-136.

Harman, H. H. *Modern Factor Analysis* (3rd ed.), University of Chicago Press, Chicago, 1976.

Harris, R. J. "Reforming Significance Testing Via Three-Valued Logic," In *What If There Were No Significance Testing?*, L. L. Harlow, S. A. Mulaik, and J. H. Steiger (eds.), Lawrence Erlbaum, Mahwah, NJ, 1997.

Harrington, S. "The Effect of Codes of Ethics and Personal Denial of Responsibility on Computer Abuse and Intention Judgments," *MIS Quarterly* (20:3), 1996, pp. 257-278.

Harrison, D. A., Mykytyn Jr., P. P., and Riemenschneider, C. K. "Executive Decisions About Adoption of Information Technology in Small Businesses: Theory and Empirical Tests," *Inforation Systems Research* (8:2), 1997, pp. 171-195.

Igbaria, M., and Guimaraes, T. "Antecedents and Consequence of Job Satisfaction Among Information Center Employees" *Journal of Management Information Systems* (9:4), 1993, pp. 145-174.

Igbaria, M., and Guimaraes, T. "Exploring Differences in Employee Turnover Intentions and Its Determinants Among Telecommuters and Non-Telecommuters," *Journal of Management Information Systems* (16:1), 1999, pp. 147-164.

Igbaria, M., Parasuraman, S., and Badaway, M. K. "Work Experiences, Job Involvement, and Quality of Work Life Among Information Systems Personnel," *MIS Quarterly* (18:2), 1994, pp. 175-201.

Jaccard, J., Turrisi, R., and Wan, C. K. *Interaction Effects in Multiple Regression*, Sage University Paper Series on Quantitative Applications in the Social Sciences 07-072, Newbury Park, CA, 1990.

Jarvenpaa, S. L., Dickson, G. W., and DeSanctis, G. "Methodological Issues in Experimental IS Research: Experiences and Recommendations," *MIS Quarterly* (9:2), 1985, pp. 141-156.

James, L., and Brett, J. "Mediators, Moderators, and Tests for Mediations," *Journal of Applied Psychology* (69:2), 1984, pp. 307-321.

Keil, M., Tan, B. C. Y., Wei, K. K., Saarinen, T., Tuunainen, V., and Wassenaar, A. "A Cross-Cultural Study on Escalation of Commitment Behavior in Software Projects," *MIS Quarterly,* 2000 (24:2), pp. 299-325.

Landis, R. S., and Dunlap, W. P. "Moderated Multiple Regression Tests Are Criterion Specific," *Organizational Research Methods* (3), 2000, pp. 254-266.

Lee, A. "Executive Overview. Validation in Information Systems Research: A State of the Art Assessment," *MIS Quarterly* (25:1), 2001, p. xi.

Lee, B., Barua, A., and Whinston, A. B. "Discovery and Representation of Causal Relationships in MIS Research: A Methodological Framework," *MIS Quarterly* (21:1), 1997, pp. 109-136.

Likert, R. A. "A Technique for the Measurement of Attitudes," *Archives of Psychology* (140), 1932.

Lubinski, D., and Humphreys, L. G. "Assessing Spurious 'Moderator Effects': Illustrated Substantively with the Hypothesized ('Synergistic') Relation between Spatial and Mathematical ability," *Psychological Bulletin* (107), 1990, pp. 385-393.

McClelland, G. H., and Judd, C. M. "Statistical Difficulties of Detecting Interactions and Moderator Effects," *Psychological Bulletin* (114), 1993, pp. 376- 390.

McKeen, J. D., and Guimaraes, T. "Successful Strategies for User Participation in Systems Development," *Journal of Management Information Systems* (14:2), 1997, pp. 133-150.

McKeen, J. D., Guimaraes, T., and Wetherbe, J. C. "The Relationship between User Participation and User Satisfaction: An Investigation of Four Contingency Factors," *MIS Quarterly* (18:4), 1994, pp. 427-451.

Nunnally, J. *Psychometric Theory*, McGraw-Hill, New York, 1967.

Podsakoff, P. M., MacKenzie, S. B., Ahearne, M., and Bommer, W. H. "Searching for a Needle in a Haystack: Trying to Identify the Illusive Moderators of Leadership Behaviors," *Journal of Management* (21), 1995, pp. 422-470.

Raelin, J. A. "Work Patterns in the Professional Lifecycle," *Journal of Occupational Psychology* (58:3), 1985, pp. 177-187.

Rice, R. E., and Contractor, N. S. "Conceptualizing Effects of Office Information Systems: A Methodology and Application for the Study of Alpha, Beta, and Gamma Changes," *Decision Science* (21), 1990, pp. 301-317.

Russell, C. J., and Bobko, P. "Moderated Regression Analysis and Likert Scales: Too Coarse for Comfort," *Journal of Applied Psychology* (77), 1992, pp. 336-342.

Russell, C. J., and Dean, M. A. "To Log or Not to Log: Bootstrap as an Alternative to Parametric Estimation of Moderation Effects in the Presence of Skewed Dependent Variables," *Organizational Research Methods* (3), 2000, pp. 167-185.

Saleem, Y, "An Empirical Test of the Contingency Approach to User Participation in Information Systems Development," *Journal of Management Information Systems* (13:1), 1996, pp. 145-166.

Saunders, D. R. "Moderator Variables in Prediction," *Educational and Psychological Measurement* (16), 1956, pp. 209-222.

Schepanski, A. "The Predictive Ability Criterion in Experimental Judgment Research in Accounting," *Decision Sciences* (14), 1983, pp. 503-512.

Schmidt, F. E. "Implications of a Measurement Problem for Expectancy Theory Research," *Organizational Behavior and Human Performance* (10), 1973, pp. 243-251.

Schmitt, N. W. "The Use of Analysis of Covariance Structures to Assess Beta and Gamma Change," *Multivariate Behavioral Research* (17), 1982, pp. 343-358.

Schmitt, N. W., and Klimoski, R. J. *Research Methods in Human Resource Management*, South-Western, Cincinnati, OH, 1991.

Schmitt, N. W., Pulakos, E. D., and Lieblein, A. "Comparison of Three Techniques to Assess Group-Level Beta and Gamma Change," *Applied Psychological Measurement* (8), 1984, pp. 249-260.

Sharma, S., Durand, R. M., and Gur-Aire, O. "Identification and Analysis of Moderator Variables," *Journal of Marketing Research* (18), 1981, August, pp. 291-300.

Stevens, S. S. "Problems and Methods of Psychophysics," *Psychological Bulletin* (55), 1958, pp. 177-196.

Taylor, S., and Todd, P. "Assessing IT Usage: The Role of Prior Experience," *MIS Quarterly* 1995 (19:4), pp. 561-570.

Thompson, R. L., Higgins, C. A., and Howell, J. M. "Influence of Experience on Personal Computer Utilization: Testing a Conceptual Model," *Journal of Management Information Systems* (11:1), 1994, pp. 167-187.

Todd, P., and Benbasat, I. "Evaluating the Impact of DSS, Cognitive Effort, and Incentives on Strategy Selection," *Information Systems Research* (10:4), 1999, pp. 356-374.

Venkatesh, V. "Determinants of Perceived Ease of Use: Integrating Control, Intrinsic Motivation, and Emotion into the Technology Acceptance Model," *Information Systems Research* (11:4), 2000, pp. 342-365.

Venkatesh, V., and Morris, M. G. "Why Don't Men Ever Stop to Ask for Directions? Gender, Social Influence, and Their Role in Technology Acceptance and Usage Behavior," *MIS Quarterly* (24:1), 2000 pp. 115-139.

Vroom, V. H. *Work and Motivation*, Wiley, New York, 1964.

Wagner, J. A. "The Organizational Tenure-Job Involvement Relationship: A Job-Career Experience Explanation," *Journal of Organizational Behavior* (8:1), 1987, pp. 63-70.

Weill, P. "The Relationship between Investment in Information Technology and Firm Performance: A Study of the Valve Manufacturing Sector," *Information Systems Research* (3:4), 1992, pp. 307-333.

Winer, B. J. *Statistical Principles in Experimental Design*, McGraw Hill, New York, 1974.

About the Authors

Traci A. Carte is an assistant professor of MIS in the Michael F. Price College of Business at the University of Oklahoma. She received her Ph.D. in MIS from the University of Georgia. Her research focuses on contingency factors related to IT implementation success. Her research has been published or is forthcoming in *Information Systems Research, MIS Quarterly, Decision Support Systems, Public Productivity and Management Review*, and numerous national and international conference proceedings.

Craig J. Russell is a professor of Management and Psychology in the Michael F. Price College of Business at the University of Oklahoma. He received his Ph.D. in business administration from the University of Iowa, majoring in human resource management and minoring in applied statistics. Craig's research focuses on advancing theory and practice in selection and development of organizational leaders. His work has appeared in *Academy of Management Journal*, *Applied Psychological Measurement, Human Resource Management Review*, *Human Relations, Journal of Applied Psychology*, *Journal of Management*, *Journal of Occupational and Organizational Psychology*, *Leadership Quarterly*, *Personnel Psychology*, and *Research in Personnel/Human Resources Management*.

Appendix A

Testing Moderated Relationships

Citation	Δr^2 Reported	Hypothesized Main Effect (No Ratio Scales)	Reported X→ Z Correlation	Coarse Scaling of Dv	Used PIs w/Sub Groups
Agarwal and Prasad (1998)	N	Y	N	Y	N
Ahituv, Igbaria, and Sella (1998)	N	N	N	N	N
Armstrong and Sambamurthy (1999)	N	Y	N	Y	Y
Banerjee, Cronan, and Jones (1998)	N	Y	N	Y	N
Banker and Slaughter (2000)	N	N	Y	N	N
Choe (1996)	Y	Y	N	Y	N
Devaraj and Kohli (2000)	N	N	Y	Y	N
Duxbury, Higgins, and Mills (1992)	N	N	N	Y	N
Fritz, Yarasimhan, and Rhee (1998)	N	N	N	Y	N
Grover, Cheon, and Teng (1996)	Y	N	Y	Y	N
Hardgrave, Wilson, and Eastman (1999)	Y	N	N	Y	N
Harrington (1996)	Y	N	N	Y	N
Harrison, Mykytyn, and Riemenschneider (1997)	Y	Y	N	Y	N
Igbaria and Guimaraes (1993)	Y	N	Y	Y	N
Igbaria and Guimaraes (1999)	N	N	N	Y	N
Igbaria, Parasuraman, and Badaway (1994)	N	Y	N	Y	Y
Keil, Tan, Wei, Saarinen, Tuunainen, and Wassenaar (2000)	N	Y	N	N	Y
McKeen and Guimaraes (1997)	N	N	Y	Y	N
McKeen, Guimaraes, and Wetherbe (1994)	Y	Y	Y	Y	N
Saleem (1996)	N	Y	N	Y	N
Taylor and Todd (1995)	N	N	N	Y	N
Thompson, Higgins, and Howell (1994)	N	Y	Y	Y	Y
Todd and Benbasat (1999)	N	Y	N	N	N
Venkatesh (2000)	N	N	N	Y	Y
Venkatesh and Morris (2000)	N	N	N	Y	Y
Weill (1992)	N	Y	N	N	N