2019

Theoretical Model Testing with Latent Variables

Robert Ping
Wright State University

Robert Ping

(HOME)

Latent Variable Research

(Displays best with Microsoft IE/Edge, Mozilla Firefox and Chrome (Windows 10)

Please note: The entire site is now under construction.

Please send me an email at rping@wright.edu if something isn't working.

FOREWORD--This website contains research on testing theoretical models (hypothesis testing) with latent variables and real world survey data (with or without interactions or quadratics).

It is intended primarily for PhD students and researchers who are just getting started with testing latent variable models using survey data. It contains, for example, suggestions for reusing one's data for a second paper to help reduce "time between papers." It also contains suggestions for finding a consistent, valid and reliable set of items (i.e., a set of items that fit the data), how to specify latent variables with only 1 or 2 indicators, how to improve Average Variance Extracted, a monograph on estimating latent variables, and selected papers.

News:

A paper about **reusing a data set** to create a second theory-test paper is available (to help reduce the "time between papers"). It turns out that an editor might not object to a paper that reuses data which has been used in a previously published paper, if the new paper's theory/model is "interesting" and materially different from the previously published paper. The paper on reusing data discusses how submodels from a previous paper might be found for a second paper that may not require collecting new data. (Please click here for more.)

Recent Additions and Changes (indicated by "New," "Revised" or "Updated"):

o Comments on specifying and estimating a manifest (i.e., observed, single-indicator, etc.) variable as a latent variable,
o suggestions for specifying and estimating a latent variable with only 2 indicators,
o comments on the use of regression in theoretical model (hypothesis) testing.
o an EXCEL template to "weed" a measure so it "fits the data" (i.e., so it is internally consistent) and
o several papers have been added, including one titled "What is Structural Equation Analysis?" that may be useful to those who would like to quickly gain a sense of this topic.
o Some of the material below is duplicated elsewhere on the web sire (click here to view the entire web site).

Please note: If you have visited this web site before, and the latest "**Updated**" date (at the top of the page) seems old, you may want to click on your browser's "Refresh" or "Reload" button on the browser toolbar (above) to view the current version of this web page.

All the material on this web site is copyrighted, but you may save it and print it out. My only request is that you please cite any material that is helpful to you. APA citations for the material below are shown with the material.

Don't forget to Refresh: Many of the links on this web site are in Microsoft WORD. If you have viewed one or more of them before, the procedure to view the latest (refreshed) version of them is tedious ("Refresh" does not work for Word documents on the web). With my apologies for the tediousness, to refresh any (and all) Word documents, please click on "Tools" on the browser toolbar (above), then click on "Internet Options...." Next, in the "General" tab, find the "Temporary Internet Files" section and click on "Delete Files...." Then, click in the "Delete all offline content" box, and click "OK." After that, close this browser window, then re-launch it so the latest versions of all the WORD documents are forced to download.

Your questions and comments are encouraged; just send an e-mail to rping@wright.edu.

Latent Variables:

Frequently Asked Question:

"**I**s there any way to speed up the process of attaining internal consistency for a measure (making a multi-item measure fit the data with more than 3 items)?" (Please see the first EXCEL template below.)

"**W**hat is structural equation analysis?"
(Please click here for a paper on this matter, then please e-mail me with any questions or comments you may have.)

"**W**hy are reviewers complaining about my use of standardized loadings?"
It turns out that standardized loadings (latent variable (LV) loadings specified
as all free so the resulting LV has a variance of unity) may produce incorrect
t-values for some parameter estimates in real world data, including structural
coefficients. This presents a problem for theory testing: An incorrect (biased)
t-value for a structural coefficient means that any interpretation of the structural
coefficient's significance or nonsignificance versus its hypothesis may be risky.
(Please click here for more.)

"**W**hy are reviewers complaining about the use of multiple regression in my paper?" (Please click here for a paper on this subject.)

"**I**s there any way to improve Average Variance Extracted (AVE) in a Latent Variable X?"
(Please click here for a paper on this matter.)

"**H**ow does one specify and estimate latent variables with only 1 or 2 indicators?"
(Please click here for a paper on this matter, then please consider e-mailing me--I have more suggestions.)

EXCEL Templates:

(CLICK ON A RED DOT)

● For Latent Variable Regression, a measurement-error-adjusted **regression approach to Structural Equation Analysis**, for situations where regression is useful (e.g., to estimate nominal/categorical variables with LV's) (see Ping 1996, *Multiv. Behav. Res.*, a revised version appears below). More about the template.

● For "weeding" a multi-item measure so it "fits the data" (i.e., **finding a set of items that "fits the data,"** so the measure is internally consistent).
Note: In real-world data, there frequently are multiple subsets of a multi-item measure that will "fit the data," and this raises the issue of which of these subsets is "best" from a validity standpoint. This template helps find at least one subset of items, usually with a maximal number of items (typically different from the one found by maximizing reliability, and, so far, containing more than 3 items), that will "fit the data." The template then can be used to search for additional subsets of items that will also fit the data, and thus it helps find the "best" face- or content valid subset of items in a measure. More about the template.

On-Line Monograph:

(CLICK ON A RED DOT)

● ***TESTING LATENT VARIABLE MODELS WITH SURVEY DATA*** (2nd Edn.)

The results of a large study of theoretical model (hypothesis) testing practices using survey data, with critical analyses, suggestions and examples. Potentially of interest to Ph.D. students and researchers who conduct or teach theoretical model testing using survey data. Contents include the six steps in theoretical model (hypothesis) testing using survey data; **Scenario Analysis**; alternatives to dropping items to attain model-to-data fit; inadmissible solutions with remedies; interactions and quadratics; and pedagogical examples (177 pp.).

Of particular interest lately is how to efficiently and effectively "weed" items to attain a consistent measure (see STEP V, PROCEDURES FOR ATTAINING...).

The APA citation for this on-line monograph is Ping, R.A. (2004). *Testing latent* *variable models with survey data, 2nd edition.* [on-line monograph]. http://www.wright.edu/~robert.ping/lv1/toc1.htm .

- Ping (2002) *TESTING LATENT VARIABLE MODELS WITH SURVEY DATA* (Edition 1)

Selected Papers on Latent Variables:

(CLICK ON A RED DOT)

- "On the Maximum of About Six Indicators per Latent Variable with Real-World Data." (An earlier version of Ping 2008, *Am. Mktng. Assoc. (Winter) Educators' Conf. Proc.*).

The paper suggests an explanation and remedies for the puzzling result that Latent Variables in theoretical model testing articles all have a maximum of about 6 indicators.

- "On Assuring Valid Measures for Theoretical Models Using Survey Data" (An earlier version of Ping 2004, *J. of Bus. Res.,* revised December 2006).

The paper reviews and comments on extant procedures for creating valid and reliable latent variable measures.

- "But what about Categorical (Nominal) Variables in Latent Variable Models?" (An earlier version of Ping 2009, *Am. Mktng. Assoc. (Summer) Educators' Conf. Proc.*).

In part because categorical variables almost always are measured in surveys in the Social Sciences (e.g., "Demographics"), the paper suggests a procedure for estimating nominal ("truly" categorical) variables in a structural equation model that also contains latent variables.

(HOME)

NOTES ON “USED DATA”--
REUSING A DATA SET TO CREATE
A SECOND THEORY-TEST PAPER

Ping R.A. (2013). "Notes on 'Used Data.'" Am. Mktng. Assoc. (Summer) Educators' Conf. Proc.

ABSTRACT

There is no published guidance for using the same data set in more than one theory-test paper. Reusing data may reduce the “time-to-publication” for a second paper and conserve funds as the “clock ticks” for an untenured faculty member. Anecdotally however, there are reviewers who may reject a theory-test paper that admits to reusing data. The paper critically discusses this matter, and provides suggestions.

INTRODUCTION

Anecdotally, there is confusion among Ph.D. students about whether or not the same data set ought to be used in more than one theory-test paper. Some believe that data should be used in only one such paper. Others believe that data may be reused.

In a small and informal survey of journal editors, none was found to be opposed to reusing data, even when their journals’ “instructions to the writers” stated or implied that the study, and presumably its data, should be original.

In an anecdote from this survey, an editor summarized his experience with a paper that used data from a previous article. One reviewer rejected the paper because the data was not “original,” while the other reviewers saw no difficulty with a paper that relied on “used data.” This anecdote hints there also may be confusion about used data among some reviewers, and, since they are likely authors, presumably among some authors.

In a small pretest of a study of faculty at Research 1 universities who had Ph.D. students, none could recall the topic of reusing data in theory tests ever being discussed.

Because the consequences of any such confusion might include that the diffusion of knowledge may be impeded (e,g., an important study could be delayed, or go unpublished, because the author(s) had difficulty funding a second study), the paper critically discusses the reuse of data in theory tests, and provides suggestions. Along the way, several matters are raised for possible future discussion and pursuit.

USED DATA

"Used data" is ubiquitous. Secondary data from, for example, the US Census Bureau, and the Bureau of Labor Statistics, are in use almost everywhere. The advantages of (re)using this data include reduced costs and time. But data collected by governments/non-governmental-organizations/commercial firms may not be ideal for a theory test. (It tends to be descriptive, and multi-item measures typical in theory tests may be unavailable; raw secondary data may be difficult to obtain; or it may not measure all the variables that are important to the researcher.)

This paper will focus on the initial reuse of primary data; typically with formative/reflective (multi item) measures intended or used for theory testing. Theory-testing situations that might be judged to involve the initial reuse(s) of data include creating two or more papers based on a single data set gathered by the author(s). Other situations include creating a paper based on data that was previously collected for commercial purposes. (Anecdotally, in Europe, Ph.D. candidates' dissertation data may have been gathered and used by a "sponsoring company" for the company's commercial purposes that are unrelated to the dissertation.) They also include reanalyzing a published data set for illustrative or pedagogical purposes (typically for a suggested methodology), and reanalyzing a paper's data to further understand or "probe" a result observed in the paper. Less obviously, improving measure psychometrics (e.g., deleting measure items to improve reliability), and model-building also involve reusing data.

The advantages and disadvantages of reusing data are discussed next. Then, suggestions for theory testing are provided, and avenues for future research are sketched.

ADVANTAGES OF RESUSING A THEORY-TEST DATA SET

One advantage of reusing data is that it can reduce the elapsed time between theory generation and analysis, the resources required for data gathering (e.g., costs), and in some cases (e.g., data gathered by others) the expertise required to gather data. For example, in a model with several variables, after a paper that tests hypothesized links among (exogenous) model antecedents and their (endogenous) consequences, more papers in which the antecedents (or the consequences) are themselves linked, might be theoretically interesting enough for submission without gathering additional data. (Criteria for “theoretically interesting” might include new theory that either extends, or fills a gap in, extant theory.)

Reusing data may enable the division of a large paper into two or more papers, in order to satisfy a journal’s page limit. For example, in a model with multiple final endogenous (consequence) variables, these variables might be divided into two sets of consequence variables (with their antecedents), and thus two papers, one for each resulting model. In each paper, this might reduce the number of hypotheses and their justifications, and the discussion and implications sections.

Stated differently, it might mean that an important study would not be delayed, or go unpublished, because of paper size, or difficulty funding an additional study.

Other advantages of reusing data might include:

- “Piggy backing” a theory test onto a commercial survey. This and using data already gathered by a commercial firm also may save time and costs.

- Combining two surveys into a single survey. Unrelated surveys may not be easily combined, but, for example, when two models have some of the same latent variables, time and money might be conserved.
- Publication of a dissertation with changes. (These changes should be based on additional theory, such as an additional path(s), that was developed prior to any data analysis beyond that for the dissertation. Stated differently, the logic of science (e.g., Hunt 1983) permits empirical discovery, hypothesis, then testing; but testing must be conducted using different data from that used in empirical discovery—see Kerr 1998 (I thank a reviewer for this citation)).
- The use of secondary data.

Although it is now less popular that it was, meta analysis (e.g., Glass 1976) uses previously gathered data. In addition, methodologists and others also have used previously published data sets to illustrate a suggested methodology (e.g., Jöreskog and Sörbom 1996, and Bentler 2006).

Reuse of a paper's data includes estimating associations "Post Hoc"--after the model has been estimated (see Friedrich 1982)--to further understand or explain an observed association(s). It also includes reanalysis of the paper's data to illustrate different model assumptions. (For example, Ping 2007 reported results with and without Organizational Commitment in the proposed model for discussion purposes.)

Reusing data also enables psychometric improvement of measures. Measure items are routinely deleted serially with measure (or model) reestimation to improve reliability and facets of validity (e.g., average extracted variance—see Fornell and Larker 1981). This might be argued to be reuse of the data set (i.e., data snooping) to find the "best" itemization of a measure.

DISADVANTAGES OF RESUSING A THEORY-TEST DATA SET

Reusing data to produce more "hits" may not be viewed others as a worthy endeavor. Absent a compelling explanation such as reducing paper size, or sharpening the focus of a paper (e.g., a previous paper was on the antecedent-consequences links, and the next paper is about the links among the consequences), a reviewer (or reader) might judge data reuse as opportunism rather than "proper" science.

A second paper that, for example, replaces correlations in a previously published model's antecedents with paths, may be judged conceptually too similar to the first paper for publication. Thus, instead of conserving time, time may be wasted on a second paper that experiences rejections because of its insufficient contribution beyond the first paper.

Further, papers that are variations on a single model, and that reuse not only data but theory/hypotheses, measures, and methods, and share some results that are identical to a previous paper could be judged idioplagaristic. As a result, time and effort may be lost in rewriting to perceptually separate papers that use the same data set.

Care must be taken in how a model is divided into submodels. For example, omitting one or more significant exogenous variables in a model may bias the path coefficients of an endogenous variable to which they are linked (i.e., the "missing variable problem"--James 1980). And, it is easy to show that omitting one or more dependent variables in a model may change model fit, and thus standard errors and model paths' significance.

"Piggy backing" onto commercial survey (or using commercial data) may save time and costs, but an academic researcher may have difficulty controlling some of the project. For example, overall questionnaire design and its testing may not be under the control of the academic researcher. Similarly the sampling frame, sampling, and activities to increase response rates also may not be under the direction of the academic researcher. Further, the appearance of an academic researcher's "independence" from the survey "issues" (i.e., the researcher is not "up to something") may be lost by not using university letterhead or return address. (Or arguably worse: using university letterhead and return address to collect data that also will be analyzed by a commercial firm). Finally, having someone else "doing some of the work" can deprive a researcher of valuable experience in data gathering. (This could be an important disadvantage: for a dissertation, demonstrating data gathering expertise is typically required.)

Last, a questionnaire that combines several surveys may be too large for its respondents: it may increase their fatigue, and it may produce echeloning, respondent irritation over similarly worded items, etc., that can increase response errors, and produce low response rates.

DISCUSSION

It may not be apparent that a model might contain candidate submodels for additional papers. Several examples might help suggest a framework for finding candidate submodels.

Finding Submodels

In Figure 1, a disguised (but actual) theoretical latent variable model (Model 1), the blank (fixed at zero) paths (e.g., A2 -> A3) could be freed to help produce submodels. To improve readability, several Model 1 latent variables were rearranged, and exogenous (antecedent) latent variables (those without an antecedent) were relabeled "A" (see Figure 3). Terminal (endogenous) consequences (latent variables that are not antecedents) were relabeled "TC," and intermediate (endogenous) latent variables were relabeled "E."

Next, each blank (fixed at zero) path was considered for being freed, then in which direction it might be freed. Then, several of these new paths were discarded because they were theoretically implausible, of little interest theoretically, or directionality could not be established (bidirectional/non recursive paths were not considered). Next, several A's were relabeled as E's.

The results included Model 1 and the (full) Figure 3 model, plus several submodels involving the A's and E's that were judged interesting enough for possible submission. For example, a submodel involving E5, and the other E's and A's (to avoid missing variable problems—A4, for example is an indirect antecedent of E5) (Submodel 1) was judged to have submission potential (E5 was judged to be an important consequence) (see Figure 4). (Submodel 1 could be abbreviated E5 = f (E4, E6, E7, Ei, Ea, Eb, A2, A4 | i = 1-3, paths among E's free as shown in Figure 3, paths among Ea, Eb, A2 and A4 free as shown in Figure 3), where "f" denotes "function of, as shown in Figure 4" and "|" means "where.")

A "hierarchy of effects" (serial) respecification of Figure 3 also was considered. Specifically, a second-order latent variable S1 was specified using Ea, A2, Eb and A4 (see Figure 2, and see Jöreskog 1971). Similarly, second-order latent variables S2 and S3 were specified using E1-E7 (see Figure 2), and the proposed sequence S1, S2, S3 then TC was specified. (Experience suggests that a second-order latent variable can be useful to combine, and thus simplify, latent variables in a model (e.g., Dwyer and Oh 1987)).

Similarly, there was an interesting submodel involving Eb (Eb = f (Ea, A2, A4)) (not shown, but see Figure 3), and another interesting submodel involving E1-E3 (Submodel 2) ({Ei} = f (A2, A4, Ea, Eb | i = 1-3, paths among A2, A4, Ea and Eb free as shown in Figure 3, paths among Ei free as shown in Figure 3), where "{}" means "set of ") (not shown, but see Figure 3). In summary, several models were found, each having a "focal consequence" latent variable(s) that was judged to be important enough to have submission potential.

Figure 6 shows a different disguised theoretical latent variable model (Model 2) where antecedent (exogenous) latent variables have been labeled "A," and terminal consequences (latent variables that are not antecedents) have been labeled "TC." In Figure 7, Model 2 was rearranged for clarity, bolded paths were added to replace the originally blank (fixed at zero) paths in Model 2, and intermediate latent variables were (re)labeled E (Model 3). Because much of the theory and many of the measures in Model 2 were new, the first paper (with Figure 6's Model 2 and no bolded paths) was too large for journal acceptance. As a result, TC3 (itself an interesting focal variable) was excised for placement in a second paper (i.e., TC3 = f(A3, Ei | i = 1-7, all paths among A3 and Ei fixed at zero) (not shown, but see Figure 7). An additional model with the focal variable E2 = f(A3, E1, E3 | bolded paths among A3, and E1 and E3 free as shown in Figure 7) (Submodel 3) was judged interesting enough for journal submission (A3 is an indirect antecedent of E2 and is specified to avoid the missing variable problem) (not shown, but see Figure 7). Another interesting model was discovered, with the bolded Figure 7 paths among E4-E7 (with A3 and E1-E3 without their bolded paths, and without TC3), that was judged to be a "hierarchy of effects" (sequential) model (i.e., first E4, next E5 or E7, then E6, then E7) (Submodel 4) (not shown, but see Figure 7).

An additional model with a theoretically plausible and interesting non-recursive (bi-directional) path between E6 and E7 (see Figure 5, and see Bagozzi 1980) also was discovered using Figure 7. (A non-recursive model that was identified—see for example Dillon and Goldstein 1984, p.447—was not immediately obvious. At least two variables were required for identification of the bi-directional path between E6 and E7: one that should significantly affect E6 but should not be linked to E7, and another that should significantly affect E7 but should not be linked to E6. Because nearly all the Figure 7 latent variables were theoretically linked to both

E6 and E7 (and could not be omitted without risking the missing variable problem), theoretically plausible demographic variables D1 and D2 were added to attain identification). Finally, a comparison of the Figure 7 model's estimates for males versus those for females was considered.

In summary, after rearranging and re-labeling the Figure 6 latent variables for clarity, previously fixed but theoretically plausible paths were freed. Then, interesting focal variables were found and submodels with as many of the Figure 6 variables as antecedents as possible (to avoid the missing variable problem) were estimated (to determine if the results were still "interesting"). In addition, the Figure 7 model was found to contain a hierarchy of effects submodel, and at least one of the paths was plausibly non-recursive. Finally, the Figure 6 model was estimated for males, then reestimated for females, and the results were compared.

Experience suggests that models with many variables may contain "interesting" submodels. Models with several "intermediate" variables (e.g., Figure 3), and those with multiple antecedents or several terminal consequences (e.g., Figure 7) also are likely to contain interesting submodels. As the examples suggested, in addition to "single consequence" submodels, linked antecedent and linked consequence submodels (e.g., Figure 7), second order, hierarchy-of-effects and non-recursive submodels are possible. Comparing model results for categories of a demographic(s) variable also might produce interesting results.

Irregularities

Unfortunately, data reuse may provide opportunities for "irregularities." For example, combining two surveys into a single survey provides an opportunity to "data snoop" across surveys. While this might generate interesting theory, it also might result in a paper that "positions" exploratory research (data snooping, then theory/hypotheses, and then a theory disconfirmation test using the data-snooped data) as confirmatory research (theory/hypotheses prior to any data analysis involving these hypotheses, then disconfirmation).

Data reuse also may provide a temptation to “position” the results of post hoc analysis as though they were originally hypothesized. For example, care must be taken that paths discovered by post hoc data analysis (e.g., to explain an hypothesized but non-significant association) are not then hypothesized as though they were not the results of data snooping.

(Parenthetically, “data snooping” also might be acceptable using a split sample, or a simulated data set. With a split sample, half of the original data set might be used for data snooping, and the other half could be used to test any resulting hypotheses. Similarly, a simulated data set might be generated using the input item-covariance matrix from the original data set, then used for data snooping. Then, the original data set could be used to test any resulting hypotheses. In both cases, the additional hypotheses, and the split half or simulated data set procedure should be mentioned in the interest of full disclosure.

Improving Psychometrics

Viewing sequentially dropping items (item weeding) to improve measure psychometrics as reanalysis of a data set, thus reusing data, may require additional discussion. Item weeding is routinely done in structural equation analysis to improve internal and external consistency, and reliability and validity in measures. These activities have been criticized (e.g., Cattell 1973; Fornell and Yi 1992; Gerbing, Hamilton and Freeman 1994; Kumar and Dillon 1987a, 1987b), however these complaints did not involve data reuse, and these objections are now seldom heard.

Item weeding is (implicitly) justified as required to separate measurement from model structure (e.g., Anderson and Gerbing 1988). (Ideally it produces a compromise between measurement model “fit” and face validity). However, it is easy to show that in real-world data these efforts can reduce the standard errors of the structural model’s path coefficients. Stated differently, item weeding could be viewed as data snooping to (perhaps inadvertently) weaken the desired disconfirmation test of a proposed model by finding itemizations that are more likely to improve the chances of “confirming” the model.

Alternatives to weeding are few. In real-world data, summing unweeded indicators may not be acceptable because the resulting measure may be unreliable. However, Gerbing and Anderson (1984) suggested in effect that deleted items could be specified as a second factor in a second-order latent variable (e.g., Jöreskog 1971). The software they suggested to expedite this task, ITAN (Gerbing and Hunter 1988) is no longer readily available, but experience suggests that in real-world data exploratory factor analysis could be used to create second-order latent variables from the “factors” (to likely reduce both the “data snooping,” and to reduce the item deletions and thus improve measure face validity).

SUGGESTIONS FOR THEORY TESTING

Authors may want to be more aware of the opportunities attending data reuse. Even if they elect not to reuse their data for publication, finding submodels might be used as way to discover additional interesting research topics. Authors could then write a second paper on an interesting submodel while conducting a new data gathering activity to test that submodel. They also might estimate the submodel using the “old” data before the new data are available, to develop at least a framework for several sections of the new paper, including possibly the reliability and validity of the submodels’ measures (these should be reconfirmed using the new data), and the results and discussion sections.

Once the new data are available, the second paper could be revised based on the new data. The used-data issue would be avoided, and time might be conserved by the parallel activities of writing a new paper while collecting data for its test.

However, given the risks that the new paper might be judged too similar to any previous paper, or it may be judged idioplagaristic, authors may elect to conserve time and funds by constructing a new paper based on the used data. In that event, the editor of any target journal probably should be contacted, to gauge their reaction to reusing data (there is the obvious matter of possibly compromising the double blind review process, even if the editor instructs the reviewers that the authors are not necessarily the same as before).

In addition, to anticipate any reviewer objections, authors should consider a "full disclosure" of the history of the data, and the paper. Specifically, any prior publication, such as publication of a previous paper involving the data, publication of the paper as an abstract, a conference paper, etc. probably should be noted to address any reviewer questions about the paper's relationship to any other published papers.

Any previous use of the data briefly should be described in the first submission of a paper that reuses data, to address any reviewer questions about the originality of the data given the sample appears to be identical to a previously published article(s). If reuse becomes an issue during review, additional details, previous paper descriptions, and assurances such as "analysis of the data for the present paper was conducted after theorizing," and "theorizing was not revised to fit the data," etc. could be provided. Further, any valid justifications, such as "the present paper is the result of pruning the prior paper to meet the page limitation," could be stated.

In addition, in a combined survey, it could be stated that extensive pretesting was conducted to reduce survey recipient fatigue; or in a study that piggybacked onto a commercial study, that the lead researcher was careful to maintain strict control of all phases of the study. Further, it could be stated that every effort was made to reduce idioplagarism, that care was taken in creating submodels to eliminate the missing variable problem, and that the model was tested with and without omitted consequent variable to estimate any bias due to model fit. (parenthetically, this “data history” also may be important after paper acceptance, so readers can gauge the acceptability of the paper for themselves).

Ideally, if data are to be reused, that decision should be made prior to any data gathering. Specifically, after the initial model is developed, any additional submodels and their hypotheses should be developed before any data are gathered. This should reduce any temptation to develop hypotheses then insert them in the original paper based on data snooping.

If the decision to reuse data is made after data has been gathered, all submodel(s) and their hypotheses should be developed before any submodel is estimated. Again, this may reduce any temptation to insert “data snooped” hypotheses in the same paper.

Addressing the matter of multiple papers with many of the same variables, and the same hypotheses for these variables, the same measures and sample, many of the same findings, etc. being judged too similar, or even idioplagaristic, may require effort. Similarity might be reduced by emphasizing that, although the new paper involves previously studied constructs, it provides important new theory about the relationships among them. For example, Submodel 1 in Figure 3 proposed previously unexplored antecedents (E4, E6 and E7) of an important variable (E5).

Reducing the appearance of idioplagarism may require writing a fresh paper, instead of rewording (or cutting and pasting), for example, the hypotheses justifications, and the descriptions of the measures, sampling, data gathering, the results, etc. of a prior paper.

Finally, if multiple papers using the same data set are jointly submitted for review, ideally each paper should acknowledge the existence of the other(s). A (brief) explanation of each could be provided, and copies might be placed on a commercial web site, for the reviewers.

Several comments may deserve emphasis: publishing similar versions of a paper, for example a conference version or an “earlier” version of a paper, could be argued to be idioplagarism. An alternative may be to consider publishing an abstract rather than a full paper. Similarly, submitting an unaltered or slightly altered paper to multiple outlets also could be viewed as idioplagaristic. (This is proscribed by many publication outlets. Typically it is discovered by having a common reviewer, and anecdotally, violation can be grounds for rejection, or desk rejection of any future submission.) One should resist the temptation to hide any reuse of data. (A reviewer who is familiar with any previous paper may question the originality of the data.)

At the risk of overdoing it, theory should always precede data analysis. Specifically, while hypotheses may be developed or revised using data, they should not be tested using the same data. (However, hypotheses developed after post hoc analysis of the data are appropriate for the paper’s discussion or future research sections--with a caveat that these results may be an artifact of the present data set, and thus are exploratory and are in need of disconfirmation in a future study.)

FUTURE RESEARCH

It may be instructive to survey Ph.D. students, journal editors, and faculty for their attitudes about reusing data. If students have either no attitude, or a weakly held one, while some journal editors and reviewers do not object, this might suggest an additional publication strategy for untenured faculty “while the P&T clock ticks.” (However, it is plausible that “top tier” journal editors and reviewers, when reviewing for these journals, might covertly object to reusing data—indeed a comment from a reviewer in the present venue hinted that they may object to reusing data.)

A similar study of these attitudes in the European Union also might be interesting. If Ph.D. students and others are encouraged, in effect, to seek a "sponsoring company" for their research (with the possibility that their academic research may become part of the sponsoring company's commercial research), this might suggest at the very least, topics for debate, if not avenues for research and publication.

SUMMARY

Because there is no published guidance concerning the use of the same data set in several theoretical model-test papers, and there may be confusion among Ph.D. students and reviewers about whether this is appropriate in theory tests, the paper critically discussed reused data in theoretical model tests, and provided suggestions.

Experience suggests that models are likely contain at least one submodel that might be a candidate for an additional paper. And, although it was anecdotal, some editors and reviewers had no objection to "used data" in theory tests. However, authors should be aware of the risks that attend used data in theory tests: reviewers may not approve of reusing data, and any subsequent paper based on used data may be judged conceptually too similar to the first paper for publication. Papers based on used data also may be judged idioplagaristic when compared to other papers to use the data. Further, care must be taken in specifying submodels to avoid the "missing variable" problem.

Suggestions for authors included that they may want to contact the editors of target journals to gauge the acceptability of a paper based on used data. And, that if data is to be reused, that decision ideally should be made prior to data collection, to reduce any temptation to add additional hypotheses to the paper based on “data snooping” the data once it was collected. And, if data are reused, authors should consider a “full disclosure” of the history of the data set.

REFERENCES

Anderson, James C. and David W. Gerbing (1988), "Structural Equation Modeling in Practice: A Review and Recommended Two-Step Approach," *Psychological Bulletin*, 103 (May), 411-23.

Bagozzi, R. P. (1980), “Performance and Satisfaction in an Industrial Sales Force: An Examination of their Antecedents and Simultaneity,” *Journal of Marketing*, 44 (Spring), 65-77.

Bentler, Peter M (2006), *EQS 6 Structural Equations Program Manual*, Encino, CA: Multivariate Software, Inc.

Cattell, R. B. (1973), *Personality and Mood by Questionnaire*, San Francisco: Jossey-Bass.

Dillon, William R. and Matthew Goldstein (1984), *Multivariate Analysis, Methods and Applications,* New York: Wiley.

Dwyer, F. Robert and Sejo Oh (1987), “Output Sector Munificence Effects on the Internal Political Economy of Marketing Channels,” *Journal of Marketing Research*, XXIV (November), 347-58.

Fornell, Claes and David F. Larker (1981), "Evaluating Structural Equation Models with Unobservable Variables and Measurement Error," *Journal of Marketing Research*, 18 (February), 39-50.

Fornell, Claes and Youjae Yi (1992), “Assumptions of the Two-Step Approach to Latent Variable Modeling,” *Sociological Methods and Research*, 20 (Spring), 291-320.

Friedrich, R. J. (1982), “In Defense of Multiplicative Terms in Multiple Regression Equations,” *American Journal of Political Science*, 26, 297-833.

Gerbing, D.W. and J. C. Anderson (1984), “On the Meaning of Within-Factor Correlated Measurement Errors,” *Journal of Consumer Research*, 11 (June), 572-50.

Gerbing, David W. and John E. Hunter (1988), *ITAN, A Statistical Package for Item Analysis*, (Available from D. W. Gerbing, School of Business, Portland State University, Portland, OR 97207).

Gerbing, David W., Janet G. Hamilton and Elizabeth B. Freeman (1994), "A Large-scale Second-order Structural Equation Model of the Influence of Management Participation on Organizational Planning Benefits," *Journal of Management*, 20, 859-85.

Glass G. V (1976), "Primary, Secondary, and Meta-Analysis of Research," *Educational Researcher*, 5 (10), 3–8.

Hunt, Shelby D. (1983), *Marketing Theory, The Philosophy of Marketing Science*, Homewood, IL: Irwin.

James, Lawrence R. (1980), "The Unmeasured Variables Problem in Path Analysis," *Journal of Applied Psychology*, 65 (4), 415-421.

Jöreskog, K. G. (1971), “Statistical Analysis of Sets of Congeneric Tests,” *Psychometrika*, 36, 109-133.

_____ and D. Sörbom (1996), *LISREL 8 User’s Reference Guide*, Chicago: Scientific Software International.

Kerr, Norbert L. (1998), “HARKing: Hypothesizing After the Results are Known,” *Personality and Social Psychology Review*, 2 (3), 196-217.

Kumar, Ajith and William R. Dillon (1987a), “The Interaction of Measurement and Structure in Simultaneous Equation Models with Unobservable Variables,” *Journal of Marketing Research*, XXIV (February), 98-105.

_____ (1987b), “Some Further Remarks on Measurement-Structure Interaction and the Unidimensionality of Constructs,” *Journal of Marketing Research*, XXIV (November), 438-444.

Ping Robert (2007), “Salesperson-Employer Relationships: Salesperson Responses to Relationship Problems and their Antecedents," Journal of Personal Selling and Sales Management, XXVII, 1 (Winter), 39-57.

Figure 1—Abbreviated Latent Variable Model (Model 1) (Disguised) (see p. 6)

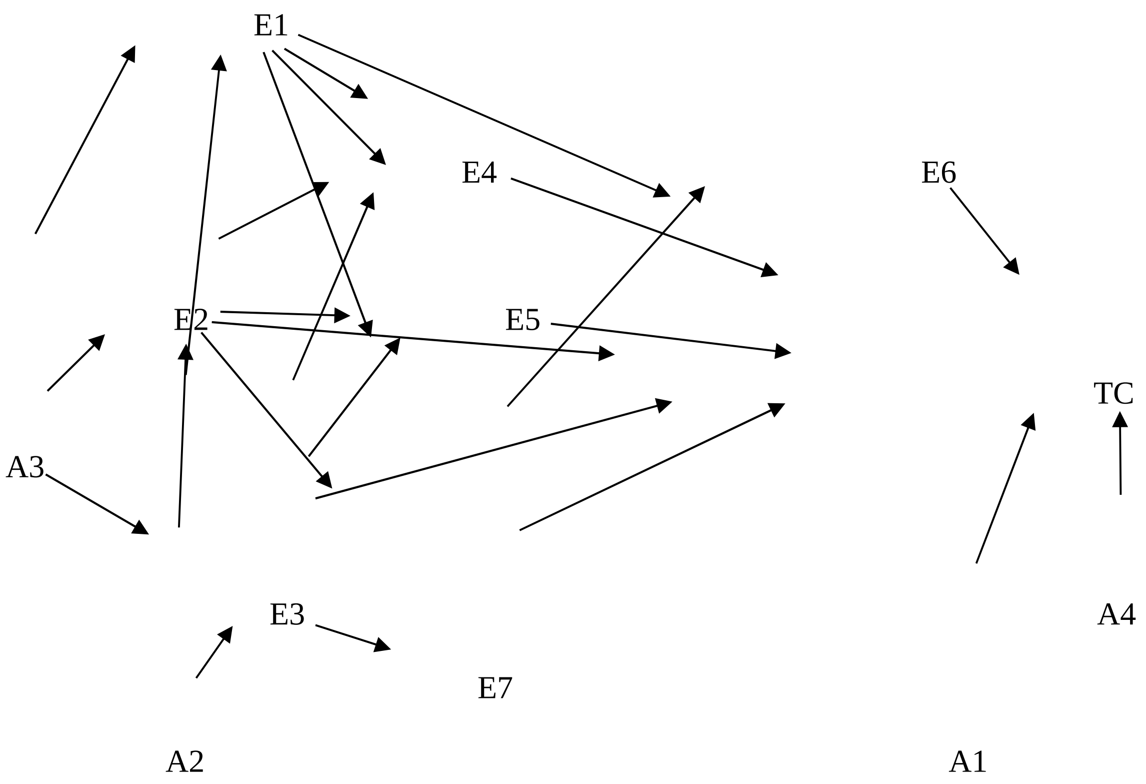


Figure 2—Respecified Figure 3 Model (see p. 7)

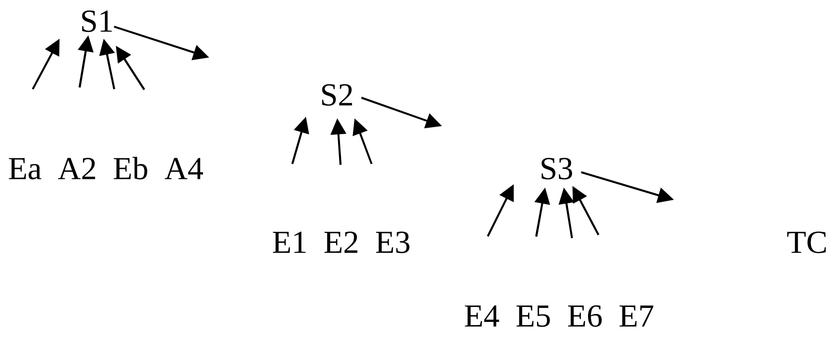


Figure 3—Rearranged Figure 1 Model with Plausible Additional Paths (in bold) (see p. 6)

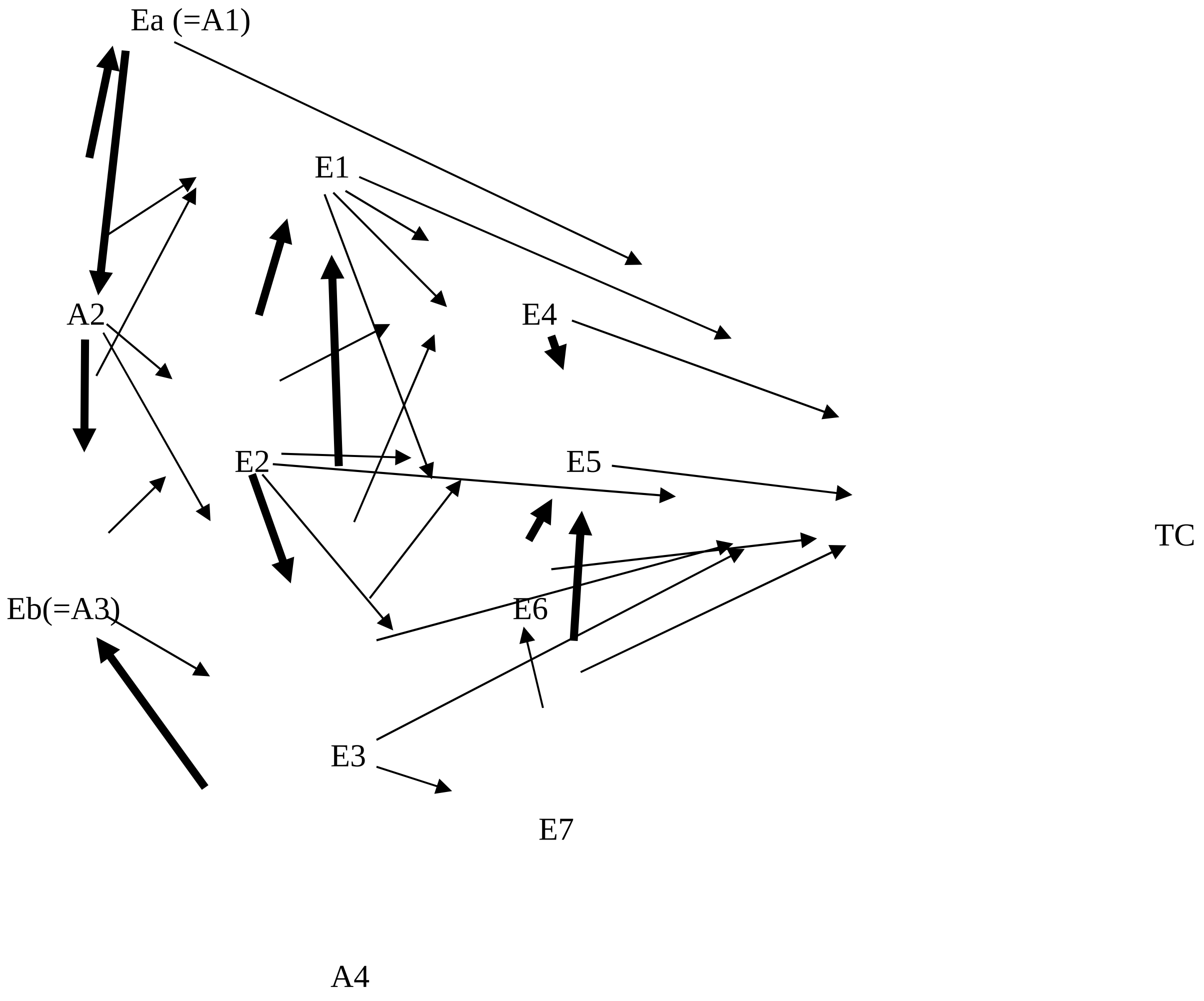


Figure 4—Submodel 1 (of Figure 3) (see p. 7)

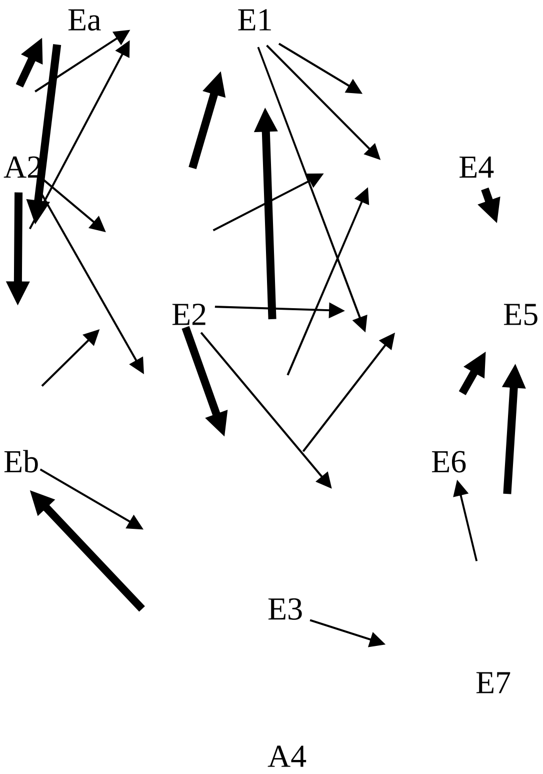


Figure 5—An Abbreviated Non-Recursive Respecification of Figure 7 (see p. 8)

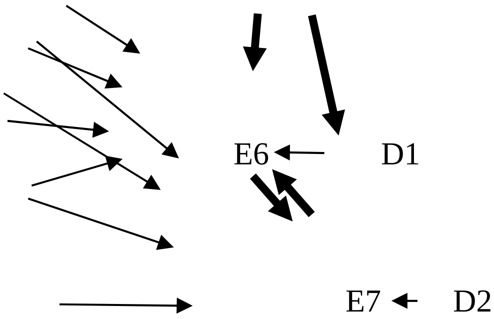


Figure 6—Abbreviated Latent Variable Model (Model 2) (Disguised) (see p. 8)

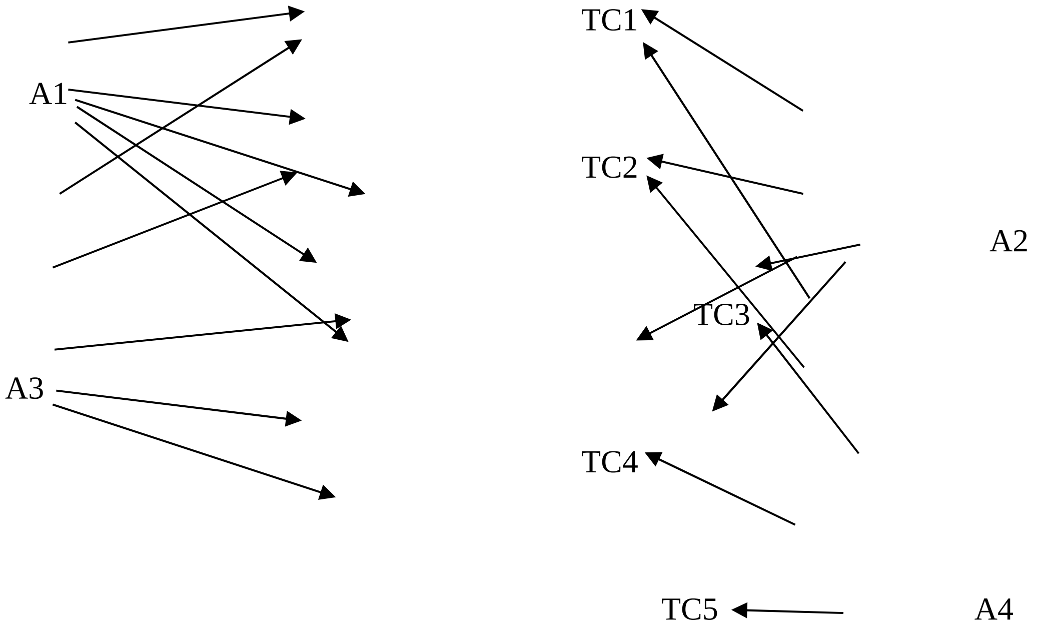
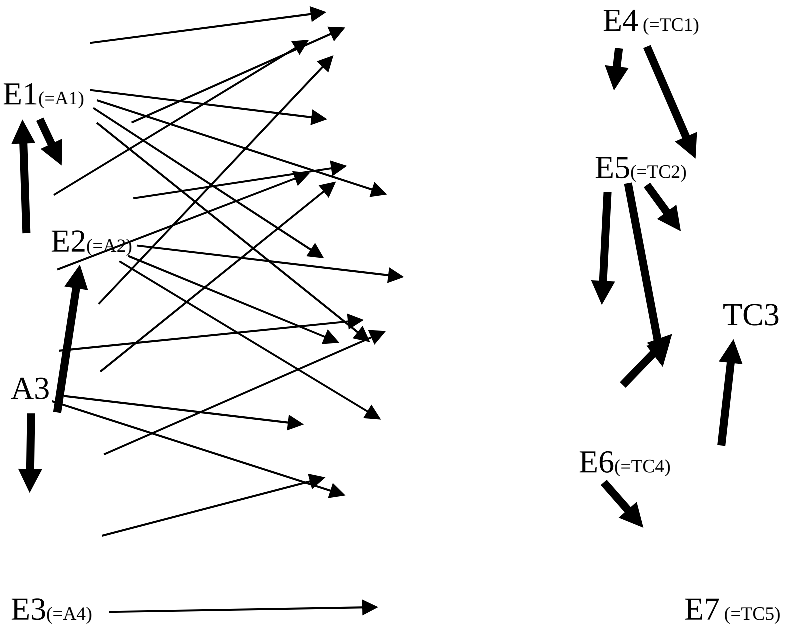


Figure 7—Rearranged Abbreviated Model 2 with Plausible Additional Paths (in bold) (see p. 8)



QUESTIONS of the MOMENT...

"What is structural equation analysis?"

(The **APA citation** for this paper is Ping, R.A. (2009). "What is structural equation analysis?" [on-line paper]. http://www.wright.edu/~robert.ping/SEA.doc)

Structural equation analysis can be understood as "regression with factor scores." In fact, even a moderate grasp of factor analysis and regression can make structural equation analysis rather easy.

Specifically, in the regression equation

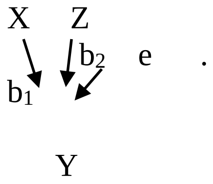
1) $Y = b_0 + b_1X + b_2Z + e$,

if X, for example, has a multiple item measure with items x_1, x_2 and x_3, sample values for X can be constructed by summing or averaging the x's in each case, and regression can proceed using the values for Z and Y in each case.

However, instead of summing or averaging the items of X, for example, factor scores could be used instead. The items of X could be factored using x_1, x_2 and x_3, and the resulting factor score for X in each case could be used in the Equation 1 regression.

Equation 1 could be diagrammed as

Figure A

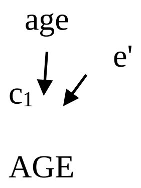


The arrows in Figure A are the plus signs in Equation 1, and the figure is read, "Y is associated with (affected by, or less commonly, "caused by") X and Z." The regression coefficients, b_1 and b_2, are shown on the arrows. Note the Equation 1 error term, e, also is diagrammed, but the intercept is not.

If X is a single-item measure such as age (e.g., How old are you?), it would have the regression equation

2) $AGE = c_0 + c_1age + e'$

that could be diagramed as



where the intercept c_0 again is missing, c_1 is a constant equal to 1 and e also is a constant equal to 0.

However, c_1 and e' need not be constants. Recalling that some respondents misreport their age, and age is usually measured on an ordinal scale that under or over-estimates each respondent's actual age, the "true score" AGE is actually a combination of the observed variable, age, and measurement error, e'. However, the (true score) AGE is unknown.

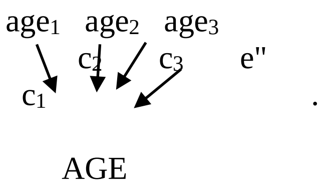
Structural equation analysis "solves" this unknown "true-score AGE problem" using three or more observations (indicators) of AGE. In this event, the items age_1 (e.g., How old are you?), age_2 (e.g., Please circle your age.) and age_3 (e.g., How old were you on your last birthday?) can be factored to produce factor scores (estimates) for the true score of AGE.

If respondents' (true score) AGE were known its regression equation would be

3) $AGE = c_0' + c_1'age_1 + c_2age_2 + c_3age_3 + e''$,

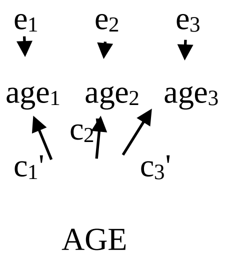
where e" is the variation of AGE not predicted by $c_1'age_1 + c_2age_2 + c_3age_3$. This could be diagramed as

Figure C



However AGE also can be "inferred" using factor analysis (i.e., from its factor scores) and its indicators age_1, age_2 and age_3, and this diagram is customarily drawn as

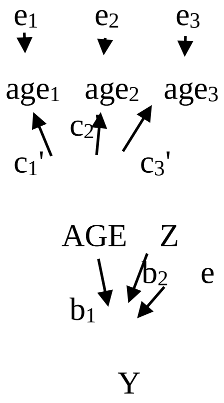
Figure D



where e" is assumed to be zero and not shown, the arrows are reversed to signal that factor analysis is involved, e_1, e_2 and e_3 the are the (measurement) errors that result when e' is assumed to be zero, and c_1', c_2' and c_3' are factor loadings (instead of regression coefficients). Stated differently, instead of regression relationships, Figure D is meant to show that the "true" (factor) score for AGE equals the observed score age_1 plus the error e_1 (i.e., $\text{AGE}_{\text{True score}} = {}^{\text{Observed}}\text{age}_1. + {}^{\text{Error}}e_1$). It also shows that the "true" score for AGE equals the observed score age_2 plus (a different) error, e_2, and the true score for AGE equals the observed score age_3 plus (another) error, e_3. (The observed scores, age_1, age_2 and age_3 are scaled by the loadings, c_1', c_2' and c_3' so the variance of AGE produced by age_1, age_2 and age_3 is the same).

Figures A and D are customarily combined in structural equation analysis as

Figure E



where X is AGE, which is "inferred" using factor analysis. Commonly seen in structural equation analysis, these diagrams are typically a combination of a regression diagram and one or more factor analysis diagrams.

Using the cases with observations (and/or factor scores) for Z and Y, the regression coefficients (the b's) in Figure E could be determined using factor scores as values for AGE.

This could be done using factor analysis and regression. Or it could be done using structural equation analysis software (AMOS, LISREL, EQS, etc.) that accomplishes the "factor-scores-for-AGE, then-regression" process by estimating Figure E in "one step." This software can be learned beginning with estimating factor scores for one factor. To illustrate, starting with an independent variable from your model, X, and its measure with the items x_1, x_2, etc., estimate factor scores for X by (exploratory) factoring it using maximum likelihood exploratory factor analysis.

Then, if X is unidimensional, estimate factor scores for X and its items x_1, x_2, etc. by factor analyzing X (alone, using only X's items, x_1, x_2, etc.) using structural equation analysis and maximum likelihood estimation. (The "diagram" that is frequently used to "program" structural equation analysis software should be similar to Figure C with X instead of AGE, and possibly more indicators for X.) The standardized factor loadings that will be available in the structural equation analysis (confirmatory) factor analysis output should be roughly the same as the factor loadings from the exploratory factor analysis. (Some will be nearly the same and a few will be considerably different, but the averages should be about the same.)

Next, using a dependent variable from your model, Y, and its measure with the items y_1, y_2, etc., exploratory factor analyze Y. If Y is unidimensional, confirmatory factor analyze the factor Y (alone), as was just done with the factor X. Again, the standardized factor loadings for Y that will be available in the structural equation analysis output should be roughly the same as the factor loadings from the exploratory factor analysis of Y.

Then, factor X and Y jointly (using X's and Y's items together) using maximum likelihood exploratory factor analysis. Next, find the factor scores for X and Y jointly by (confirmatory) factor analyzing X and Y using structural equation analysis (allow X and Y to be correlated). As before, the factor loadings for X in the joint exploratory factor analysis of X and Y should be roughly the same as X's standardized factor loadings shown in the joint structural equation analysis output. The same should be true for Y. The joint loadings for X should be nearly identical to those from factor analyzing X by itself (i.e., the loadings of X should be practically invariant across measurement models). Similarly, the joint loadings for Y should be practically invariant when compared to the measurement models for Y by itself.

Finally, regress Y on X using averaged indicators. Then, replace the correlated path between X and Y in the joint (confirmatory) factor analysis of X and Y above with a directional path from X to Y to produce the structural model of X and Y. The regression coefficient for X should be roughly the same as the directional path (structural) coefficient from X to Y. Further, the loadings in the structural model should be only trivially different from those in the measurement models above.

The balance of your model now could be added one unidimensional variable at a time to produce a series of exploratory factor analyses, confirmatory factor analyses, a regression, and the full structural model. As before, loadings for each factor (latent variable) should be roughly the same between the exploratory and confirmatory factor analyses, and the structural equation analysis loadings for each latent variable should be practically invariant. (The regression and structural coefficients will change as more variables are added, but corresponding latent variables should have roughly the same regression and structural coefficients.)

If loadings are not "trivially different" as more latent variables (e.g., Z) are added, this suggests that the unidimensional measures are not unidimensional enough for structural equation analysis. To remedy this, examine the Root Mean Error of Approximation (RMSEA) in the measurement models for the latent variables added so far. Improve reliability for the latent variable(s) in their (alone) measurement model(s) with an individual RMSEA that is more than .08 by deleting items that reduce its reliability. (Procedures for this are available in SAS, SPSS, etc.) Do this for any other latent variable that has an (alone) RMSEA more than .8 as it is added. The result should be that corresponding latent variable loadings are practically invariant across all their measurement and structural models.

After all the latent variables have been added to the model, the results are that the model's structural coefficients have been estimated using structural equation analysis, and the full measurement and structural models (i.e., with all the variables) "fit the data" (their RMSEA's are less than .08).

If the structural model does not fit the data, but the full measurement model does, it is because there is a path somewhere that should not be assumed to be zero. Because this exact path could be anywhere, structural equation analysis works best on models in which all paths are (adequately) predicted by theory.

Obviously there is more to learn. But, estimating your model's structural coefficients was the immediate objective, and one could use the above process again to estimate other structural equation models.

However, you may be wondering why factor scores were never actually used. Structural equation analysis can produce factor scores, but they are not used in the actual structural equation analysis computer algorithm.[1] Factor scores were a pedagogical device used to help explain things.

You also may be wondering, if regression estimates are "roughly the same" as structural equation analysis estimates, why not use the regression estimates (or exploratory factor analysis factor scores with regression)? The problem is that regression "sameness" becomes "rougher and rougher" as more latent variables are added to the model (the direction (sign) of one or more coefficients can eventually be different between regression and structural equation analysis).

[1] Structural equation analysis minimizes the difference between the input covariance matrix of the observed items and the covariance matrix of these items implied by the measurement or structural model. For example, in the Figure E model the indicators of AGE and the variable Z were not allowed to correlated (i.e., they are assumed to have no paths connecting them), even though their input data are correlated.

Simulation studies have "confirmed" (consistently suggested) that with multi-item measures, known factor structures, with known loadings and known regression coefficients, are "better" estimated by structural equation analysis' "minimize the (chi-square) difference between the input covariance matrix of the observed items and the covariance matrix of these items implied by the measurement or structural model" than by regression. For this reason, regression estimates now labeled as "biased" when one or more multi-item measure is present in a regression equation (e.g., Aiken and West 1991, Bohrnsted and Carter 1971, Cohen and Cohen 1983, Kenny 1979).

Some of the structural equation analysis jargon and "standard practice" may be of interest.

> There is little agreement measures of model to data fit (e.g., Bollen and Long 1993). I used RMSEA because it is adequate for these purposes.
>
> In the measurement and structural models, one indicator of each latent variable is customarily "fixed" (set) to the value of 1.
>
> Reliability and validity receive considerable attention in structural equation analysis. However, there is little agreement on validity criteria (see **QUESTIONS of the MOMENT**... "What is the "validity" of a Latent Variable Interaction (or Quadratic)?" for details).
>
> Improving model-to-data fit by maximizing reliability can degrade construct or face validity (again see **QUESTIONS of the MOMENT**... "What is the "validity" of... "). In general, care must be taken when deleting items from a measure in the measurement or structural model estimation process.
>
> In structural equation analysis, significance is customarily suggested by a t-value greater than 2 in absolute value (p-values are not used).
>
> Structural equation analysis can accommodate multiple dependent variables in a single model. And, dependent variables can affect each other. For these reasons a dependent variable is termed an endogenous variable in structural equation analysis (independent variable are called exogenous variables).
>
> In structural equation analysis models it is standard practice to correlate (free) exogenous variables, but not to correlate endogenous variables.
>
> The term "consistency" is used to imply the stronger unidimensionality required by structural equation analysis (e.g., the indicators of X are consistent).

References

Aiken, L. S. and S. G. West (1991), *Multiple Regression: Testing and Interpreting Interactions*, Newbury Park, CA: Sage.

Bohrnstedt, G. W. and T. M. Carter (1971), "Robustness in regression analysis," in H.L. Costner (Ed.), *Sociological Methodology* (pp. 118-146), San Francisco: Jossey-Bass.

Bollen, Kenneth A. and J. Scott Long (1993), *Testing Structural Equation Models*, Newbury Park, CA: SAGE Publications.

Cohen, Jacob and Patricia Cohen (1983), *Applied Multiple Regression/Correlation Analyses for the Behavioral Sciences*, Hillsdale, NJ: Lawrence Erlbaum.

Kenny, David (1979), *Correlation and Causality*, New York: Wiley.

QUESTIONS of the MOMENT...

"Why are reviewers complaining about my use of standardized loadings?"

(The **APA citation** for this paper is Ping, R.A. (2013). "Why are reviewers complaining about my use of standardized loadings?" [on-line paper]. http://www.wright.edu/~robert.ping/stdLoad.doc)

Jöreskog (1996, "LISREL 8 … Reference Guide, p.35) warned that standard errors (in LISREL), among other statistics, may be incorrect when correlations are analyzed (without standard deviations, etc.) in structural equation models. This presents a problem in theory testing—an incorrect (biased) standard error for a structural coefficient means that its t-value is incorrect, and any interpretation of the observed structural coefficient's significance or nonsignificance versus its hypothesis may be risky.

While I have yet to find equivalent warnings about correlations and standard errors in documentation for EQS or AMOS, other authors have warned against analyzing correlations (see the citations in Bentler 2006, "EQS 6 … Program Manual," p. 11). As a result, it may be prudent to avoid analyzing correlations in theory tests involving structural equation analysis.

However, it is easy to show using real-world data that covariances and "standardized loadings" (latent variable (LV) loadings specified as all free—so the resulting LV has a variance of 1) may produce incorrect t-values for parameter estimates, including structural coefficients. Specifically, the t-values of the resulting structural coefficients (which are now standardized estimates) may be different from those produced by the preferred "unstandardized loadings" LV specification, where one loading of each LV is fixed at 1, and each LV's estimated (error-disattenuated) variance is different from 1 (e.g., Jöreskog 1996).

Thus, it also may be prudent to avoid using standardized loadings in theoretical model tests involving structural equation analysis. (If standardized coefficient estimates are required, standardized and unstandardized estimates could be requested, and standardized values could be reported with unstandardized t-values.)

Parenthetically, one procedure for specifying unstandardized LV loadings is to specify each LV with its first indicator fixed at 1. To simplify any subsequent interpretation of loadings, I then respecify each LV by fixing the largest loading of each LV to 1, and freeing each LV's first indicator if it is not the largest (to avoid having two indicators fixed at 1), and reestimate the model.

QUESTIONS of the MOMENT...

"Why are reviewers complaining about the use of multiple regression in my paper?"

(The **APA citation** for this paper is Ping, R.A. (2009). " Why are reviewers complaining about the use of multiple regression in my paper?" [on-line paper]. http://www.wright.edu/~robert.ping/MR.doc)

Multiple regression assumes each independent variable is measured without error (i.e., the observed score is exactly the true score). Unfortunately, it is well known that the extent and direction of all regression coefficient is biased by even a single variable that contains (known or unknown) measurement error (e.g., Aiken and West 1991, Bohrnsted and Carter 1971, Cohen and Cohen 1983, Kenny 1979).

Even though this assumption was well known, it was routinely ignored in theoretical model (hypothesis) testing until Jöreskog's proposal that, among other things, allowed modeling of measurement error (Jöreskog 1970, 1971) (i.e., structural equation analysis). As a result, reviewers may reject substantive papers that rely on regression because 1) its regression's assumption of variables with no measurement error is now believed to be violated even in demographic variables such as age and income (both are typically misreported by some respondent groups, and each is typically measured in "round numbers"). 2) reviewers are (re)aware of how regression estimates can be biased (i.e., untrustworthy) in theoretical model tests when one or more variable contains measurement error (unless they are uncorrelated with any of the other independent variables, which is unlikely in real-world data). And, 3) regression usually produces Least Squares estimates--Maximum Likelihood estimates are now preferred for theoretical model testing.

As a result, some reviewers now believe that regression is an insufficient test of a theoretical model if there is measurement error in even one model variable (i.e., all the resulting coefficients used to test the hypotheses are untrustworthy).

Many suggested procedures for multiple regression (e.g., Cohen and Cohen 1983) are now considered inappropriate for theory testing because for example, the analysis procedures (e.g., stepping variables in, etc.) also are insufficient tests of the hypotheses.

Alternatives to ordinary least squares regression that account for measurement error include Fuller (1991) and Ping (1996), but each has drawbacks. Fuller's proposals are inaccessible to many substantive researchers. Ping's proposal relies on measurement parameter estimates from structural equation analysis, and begs the question, why not just use structural equation analysis?

The "problems" with utilizing the now preferred structural equation analysis, appear to be several: it is not taught in all terminal degree programs. And, despite texts apparently aimed at "self teaching" it (e.g., Byrne 1990), and (powerful) graphical user interfaces now available in most structural equation analysis software packages, anecdotally, structural equation analysis still seems to be inaccessible to many substantive researchers when compared to regression. For untenured researchers who may be "on a clock," this can slow productivity. For others, this can require "finding" someone who does structural equation analysis, then "managing" their involvement in the resulting paper. Structural equation analysis also can appear to "take over" a theoretical piece, producing a perhaps unwelcome intrusion on its theoretical matters.

"Solutions" to the structural equation analysis "problems" all have drawbacks. First, if structural equation analysis is not required (e.g., for a dissertation), to conserve time don't use it. However, for the reasons stated above, this may be a temporary solution.

Next, consider allowing about a month to do three things: first, finding someone to help with learning structural equation analysis, then learning only enough structural equation analysis to "get by" reviewers. Then, consider quickly creating/revising a paper with a simple model (or a simple submodel of your current model) that uses (replaces regression with) structural equation analysis, and submitting it to a good conference. Rather than acceptance, the objective would be to learn structural equation analysis in a realistic setting. Any reviewer feedback would also suggest what/where more structural equation analysis work is needed.

Click here for more about structural equation analysis as "regression using factor scores instead of averaged items," and how to learn the basics in a reasonable amount of time.

References

Aiken, L. S. and S. G. West (1991), *Multiple Regression: Testing and Interpreting Interactions*, Newbury Park, CA: Sage.

Bohrnstedt, G. W. and T. M. Carter (1971), "Robustness in regression analysis," in H.L. Costner (Ed.), *Sociological Methodology* (pp. 118-146), San Francisco: Jossey-Bass.

Byrne, B.M. (1990). *A Primer of LISREL: Basic Applications and Programming for Confirmatory Factor Analytic Models*. New York: Springer-Verlag Inc.

Cohen, Jacob and Patricia Cohen (1983), *Applied Multiple Regression/Correlation Analyses for the Behavioral Sciences*, Hillsdale, NJ: Lawrence Erlbaum.

Fuller, Wayne A. (1991), "Regression Estimation in the Presence of Measurement Error, " in *Measurement Errors in Surveys*, B. P. Beimer, R. M. Groves, L. E. Lyberg, N. A. Mathiowetz and S. Sudman, eds., NY: Wiley.

Jöreskog, Karl G. (1970), "A General Method for Analysis of Covariance Structures," *Biometrika*, 57, 239-251.

________ (1971) "Simultaneous Factor Analysis in Several Populations," *Psychometrika*, 57, 409-426.

Kenny, David (1979), *Correlation and Causality*, New York: Wiley.

Ping, R.A. (1996c), "Latent Variable Regression: A Technique for Estimating Interaction and Quadratic Coefficients," *Multivariate Behavioral Research*, 31 (1), 95-120.

QUESTIONS of the MOMENT...

"Is there any way to improve Average Variance Extracted (AVE) in a Latent Variable (LV) X?"

(The APA citation for this paper is Ping, R. A. (2009). "Is there any way to improve Average Variance Extracted (AVE) in a Latent Variable (LV) X (Revised)?" [on-line paper]. http://www.wright.edu/~robert.ping/ImprovAVE2.doc)

(Click here for an earlier version of this paper, Ping, R. A. (2007). "Is there any way to improve Average Variance Extracted (AVE) in a Latent Variable (LV) X?" [on-line paper]. http://www.wright.edu/~robert.ping/LowAVE.doc.)

Average variance extracted (AVE) almost always can be improved by dropping cases, or by dropping the item with the largest measurement error variance. The result may be desirable to improve AVE, or to raise it above the square of a correlation with another latent variable (LV) (i.e., to improve discriminant validity in the Fornell and Larker (1981) sense).

One approach to dropping cases is to use a "Jacknife-like" procedure (Efron 1981). Specifically, a case is removed from the data set, and AVE is computed for the remaining cases.[1] Then, the removed case is replaced, a different case is removed, and AVE is computed for the remaining cases. This process is repeated for each of the rest of the cases to find the case that produces the largest AVE improvement.[2]

Additional AVE improvement may be obtained by repeating this process using the improved AVE data set (i.e., with the case that produces the largest AVE improvement removed), instead of the full data set. Specifically, a case is removed from the first improved AVE data set, and AVE is computed for the remaining cases. Then, the case just removed (not both cases) is replaced, a different case is removed, and AVE is computed for the remaining cases. This process is repeated for each of the rest of the cases to find the largest AVE improvement with two cases removed (but, see Footnote 2).

This process could be repeated using combinations of the above and Footnote 2) procedures, but experience suggests that dropping about three cases or approximately a .05 AVE improvement is the most AVE improvement that dropping cases will produce in real-world data.

[1] In a (maximum likelihood) exploratory factor analysis, the "Percent of Variance Explained" by Factor 1 can be used to gauge changes in AVE in a unidimensional set of items--it is roughly the same as AVE.

[2] Dropping the case that detracts most from AVE is arguably "not random," and this casts a shadow over sample "representativeness." An improvement would be to randomly select a case from the set of cases that detract most from AVE. Alternatively, cases could be deleted randomly and the first case that improves AVE could be dropped.

Dropping an item also will improve AVE, frequently by more that deleting cases will. However, the procedure for this is "messy," and the resulting set of higher AVE items can be less content or face valid than before items were dropped (i.e., the resulting set of items may match its conceptual and/or operational definitions less well). The results also may be less internally consistent (i.e., the single construct measurement model of the resulting items may fit the data less well).

Experience suggests that in real-world data, several consistent subsets of items from a measure usually can be found. An alternative to dropping items is to create one or more additional subsets of items and gauge the AVE of each these subsets. Replacing any deleted cases, (maximum likelihood) exploratory factor (with varimax rotation) the full measure. Then for the Factor 1 items, find highest reliability subset of these items (there are procedures in SPSS, SAS, etc. to accomplish this, usually in the "reliability" procedures). If the AVE of the resulting highest reliability subset of items is unacceptable, try dropping cases from this subset of items. (Dropping the item with the largest error term from the highest reliability subset of items usually reduces AVE.)

If model-to-data fit or content validity problems arise when items are dropped, combinations of one or more of the following procedures could be used.

Replacing any deleted cases, consider finding another consistent subset of items using Modification Indices (see Appendix A--Item Weeding in "On the Maximum of About Six Indicators..." on this web site). If this subset has unacceptable AVE, try dropping the item with the largest error, and/or drop case(s). Experience suggests that Modification Indices sometimes works better than maximizing reliability. However, AVE improvement seems to be limited to about 10 points (i.e., 0.10).

If AVE is still unacceptable, replace any deleted cases, then, using the Factor 1 items, drop the item with the largest error term first. Then, reitemize the resulting set of items using Modification Indices or by maximizing reliability. Next, drop cases, and/or drop the item with the largest error from the results.

However, if AVE of the resulting measure is within a few points of "acceptable" (0.50), this may not always be "fatal" to publishing a model test. Experience suggests that not all reviewers accept AVE as "the" measure of convergent validity, some prefer reliability. Thus, if an LV is reliable, that may be a sufficient demonstration of convergent validity for some reviewers.

In addition, the logic for possibly ignoring low AVE might be that many "interesting" theoretical model-testing studies involve a "first-time" model, and an initial model test, that together should be viewed as largely "exploratory." This "first test" usually uses new measures in a new model tested for the first time, etc., and insisting that the new measures be "perfect" may be inappropriate because new knowledge would go unpublished until a "perfect" study is attained. AVE adherents of course might reply that concluding anything from measures that are more than 50% error is ill advised, because there are so few replication studies.

In my opinion, an AVE slightly below 0.50 might be acceptable in a really "interesting" "first-time" study, 1) if it does not produce major discriminant validity problems (discussed below), 2) the diminished AVE is noted and discussed in the Limitations section of the paper, 3) any significant effects involving the low AVE LV's are held to a higher significance requirement (e.g., |t| >= 2.2 rather than |t| >= 2.0), and 4) any discussion of interpretation, and especially implications, involving the low AVE LV's are clearly labeled as "very provisional" and in need of replication.

Again, the logic would be that the model may be too interesting to suppress its first test. In different words, the focus of the paper should be on the new theory developed, and the contributions include a "first test," and that more measurement work is needed on the low AVE measures. (A less desirable alternative with low AVE would be a propositional paper, which might be considerably less "interesting.")

This "first-time study" argument also may apply when there are discriminant validity problems, and more measurement work is needed on the low discriminant validity measures. Nevertheless, it always possible to reduce the correlation between two LV's using a procedure similar to Residual Centering (see Lance 1988). The procedure involves reducing the covariation between the target LV's X and Z until the squared correlation between them is less than the AVE of both X and Z. Specifically, average the indicators for X and Z, then regress the lower AVE LV, X for example, on the higher AVE LV, Z, to produce Z = $b_0 + b_1X$. Next, subtract a percentage of $b_0 + b_1X$ (i.e.,

1) $K*(b_0 + b_1X)$,

where K is between 0 and 1) from Z in each case to scale (reduce) the covariance (and thus the (squared) correlation) between X and Z.

However experience suggests that in real-world data, scaling simply masks a discriminant validity problem rather than remedying it. Specifically, in real-world data, experience suggests that with lower AVE and correlated X and Z, the unique error variance (i.e., error variance that is unshared in the correlation between X and Z) of one or both X and Z can be greater than 50%. This in turn increases the instability (variability) of structural coefficients involving X or Z across studies beyond that which could be expected with sampling variation. Stated differently, with declined AVE and (even moderately) correlated X and Z, their association with Y, for example, can be largely the result of measurement error, which should produce different results (i.e., instability--reduced "reproducibility" in Campbell and Fiske's (1959) terms) in subsequent studies. Increased correlation between X and Z, especially when it is greater than X or Z's AVE, increases this potential for instability.

1) As a perhaps surprising example from a real-world survey, the AVE's of two LV's were both 0.59, and their correlation was -0.59 (their covariance, the square of their correlation, was 0.33). Scaling one LV to zero correlation with the other, reduced its AVE to 0.47, and scaling the other LV to zero (after removing the previous scaling)

reduced its AVE to 0.48. Thus, the amount of unique error variance in the LV's was 53% (= 1 - 0.47) in one, and 52% (1 - 0.48) in the other. This suggests that any associations involving X or Z and a dependent variable is (slightly) more the result of error variance than it is the result of error-free variance, which should (slightly) amplify any difference in results from sampling variation in subsequent studies. Note that both LV's were discriminant valid using Fornell and Larker's (1981) AVE's-versus-squared-correlation discriminant validity criterion, and they likely would not have had their discriminant validity questioned--both LV's had "acceptable" AVE's, and their correlation was less than |0.70|. Also note that in this case the unacceptable unique err-free variances might be remedied by increasing AVE in the LV's.

As another example again using real-world data, the AVE's of two LV's were 0.58 and 0.72, while their correlation was .82 (their covariance, the square of their correlation, was 0.67). Scaling the larger AVE LV to zero correlation, reduced its AVE to 0.44. and scaling other LV (after undoing the previous scaling) reduced its AVE to 0.34. In different words, the amount of unique error variance in the larger AVE LV was 56% (= 1 - 0.44), and the amount of unique error variance in the smaller AVE LV was 66% (= 1 - 0.34). This suggests that their associations with another LV,Y for example, are more the result of error variance than error-free variance, which should produce more instability (different results) in subsequent studies than if it were lower.

Finally, the AVE's of two other LV's were 0.72 and 0.87, while their correlation was -.71 (their covariance, the square of their correlation, was 0.50). Scaling the larger AVE LV to zero correlation, reduced its AVE to 0.77, and scaling other LV (after undoing the previous scaling) reduced its AVE to 0.55. In different words, the amount of unique error variance in the larger AVE LV was 23% (= 1 - 0.77), and the amount of unique error variance in the smaller AVE LV was 45% (= 1 - 0.55). This suggests that their associations with another LV,Y for example, are more result of error-free variance than error variance, which should produce (comparatively) less instability (differing results) in subsequent studies than if error variance were higher. Note that both LV's had "acceptable" AVE's, but their correlation was slightly greater than |0.70|.

These examples suggest that Fornell and Larker's AVE's-versus-squared-correlation (discriminant validity) test may or may not signal a problem with unique error variance, and thus Fornell and Larker's discriminant validity test may or may not signal declined "reproducibility," Campbell and Fiske's stated objective of validity.

Several comments may be of interest. As the first example suggests, low AVE's should be investigated for low unique error-free variance. There probably can be no firm rule, but Fornell and Larker's "AVE at least 0.50" may be insufficient. Experience suggests that all correlations above 0.7 should be investigated (see Example 1 above), especially when the AVE's of the LV's involved are less than 0.6.

Any low unique error-free variance problems should be discussed in the Limitations section of the study's paper, and any discussion of the implications of the associations involved should be prefaced with a caveat that these associations are mostly error and may be an artifact of the study.

Z and its t-value is unchanged by scaling (i.e., subtracting K*(b_0 + b_1X) from Z in each case). However, scaling reduces the variance of Z, and thus it reduces any standardized structural coefficient (beta) involving Z. The range of Z is also reduced by scaling.

For an LV, X, that fails Fornell and Larker's AVE's-versus-squared-correlation (discriminant validity) test with Z, it is easy to show that all of X's error free variance is not contained in the covariance of X and Z. (X's AVE less than its correlation with Z might mean that all of X's error-free variance, its AVE, is contained in the covariance.)

For example when two LV's, with AVE's of 0.49 and 0.72, and a squared correlation of 0.65 (i.e., their covariance (squared correlation) was larger than one LV's error-free variance (AVE), a failure of Fornell and Larker's discriminant validity test), had the smaller AVE LV's variance scaled to zero correlation between them (i.e., K = 1 in Equation 1), in a measurement model of the resulting smaller AVE LV, it had 21% error free variance (i.e., a 0.21 AVE). In different words, all the covariation between the LV's was removed by scaling, yet there still was error-free variance (AVE) in both LV's (i.e., 21% error-free variance in the smaller AVE LV and 72% in the larger).

This could be interpreted as suggesting that the LV's were operationally distinct (the customary meaning of discriminant validity[3]). (In real-world data, only if the variance of an LV is equal to its covariance with another LV is there complete operational indistinctness--this matter is further discussed below).

In real-world data, experience suggests that improving an LV's AVE does not materially change correlations with that LV.

Substantive authors have used other single-sample discriminant validity tests besides Fornell and Larker's AVE's-versus-squared-correlation (discriminant validity) test, and these tests may be attractive when there are problems with discriminant validity. These tests include testing the correlation confidence interval (see Anderson and Gerbing 1988) or a single degree of freedom test (see Bagozzi and Phillips 1982). However, it is easy to show that these tests are likely to produce untrustworthy results in theory tests with survey data. Specifically, in theory tests with real-world survey data, testing the correlation confidence interval for two LV's to see if it contains 1, which would suggest that the two LV's are empirically (operationally) indistinct (i.e., they are "discriminant invalid" in the popular sense--see Footnote 3), almost always suggests empirical distinctness. Typical sample sizes (hundreds of cases) and the internal consistency requirement in survey data theory tests typically combine to produce small correlation

[3] A careful reading of Campbell's writings suggests that their notion of discriminant validity is evidenced by low correlations with other variables, rather than the popular requirement of a lack of population correlations of 1 (i.e., empirical or operational distinctness).

standard errors that in turn produce confidence intervals are usually too small to include a correlation value of 1.

For example, two LV's that were known to be theoretically indistinct (their items contained only slight variations in item wording) and that had a correlation of 0.9988, produced a 95% correlation confidence interval of [.9984, .9993] in a sample of 200 surveys. In different words, the 95% confidence interval for these LV's suggested they were operationally distinct (i.e., "discriminant valid") even though they were theoretically indistinct and had a correlation of .9988.

A single degree of freedom test can be applied to a measurement model containing the two target LV's, or it can be applied to the full measurement model containing the target LV's, to compare the model to one where the two LV's correlation is constrained to 1. In a two-LV measurement model, a single degree of freedom test frequently produces untrustworthy test results (i.e., two highly correlated LV's that have their correlation constrained to 1 will usually fit the data significantly worse that when their correlation is free). For example, in a two-LV measurement model, two LV's that may or may not have been theoretically distinct (they were two factors of the same LV) had a correlation of .9295. In a single degree of freedom test, their correlation could not be constrained to 1 in LISREL (the fitted covariance matrix was not positive definite). However, constraining the correlation to 0.9795 instead of 1 produced a chi square difference (chi square = 388 for the constrained correlation model, chi square = 207 for the unconstrained correlation model) with 1 degree of freedom (degrees of freedom = 54 for the constrained correlation model, degrees of freedom = 53 for the unconstrained correlation model) that was significant (chi square difference = 388 - 207 = 181, which has 1 - α of 1.0000 with 1 degree of freedom), suggesting they were (very) operationally/empirically distinct (and thus "discriminant valid").

Then, two sets of items, that could be argued to be conceptually the same (both were from Factor 1 of the same LV), with a correlation of 0.9969, were tested in a two-LV measurement model. Again in a single degree of freedom test, their correlation could not be constrained to 1 in LISREL. However, constraining the correlation to 0.9998 instead of 1 produced a chi square difference (218 = 249 for the constrained correlation model, minus 31 for the unconstrained model) with 1 degree of freedom (9 for the constrained correlation model, degrees of freedom = 8 for the unconstrained model) that was significant (1 - α = 1.0000 with 1 degree of freedom), which suggested they were (very) operationally distinct, and thus "discriminant valid."

However, the chi square statistic is sensitive to sample size. So, the sample size was reduced in steps until the parameter estimates became unstable, comparing chi square differences at each step. Nevertheless, the chi square difference tests continued to be significant, suggesting that sample size did not affect these results.

A full measurement model produced similar results. Two LV's that may or may not have been conceptually the same (they were two factors of the same LV) had a correlation of .9274. While their correlation could not be constrained to 1 in a larger measurement model containing them, constraining them to a correlation of .9476 produced a chi square difference test with a significance of .9990. This suggested they were operationally distinct (i.e., they were "discriminant valid").

Then, two sets of items, that could be argued to be conceptually indistinct (both were from Factor 1 of the same LV), with a correlation of 0.9998, were tested in a two LV measurement model. In this case their correlation could be constrained to 1, and the chi square difference (= 12 = (5626 for the constrained correlation model, minus 5626 for the unconstrained model)) with 1 degree of freedom (1836 for the constrained correlation model, 1835 for the unconstrained model) was significant (1 - α = .9999 with 1 degree of freedom), which suggested they were (very) operationally distinct, and thus "discriminant valid."

Again, because the chi square statistic is sensitive to sample size, the sample size was reduced in steps until the parameter estimates became unstable, comparing chi square differences at each step. Again, the chi square difference tests continued to be significant, suggesting that sample size did not affect these results.

In summary, experience suggests that in real-world data, alternative "discriminant validity" tests, such as correlation confidence intervals or single degree of freedom tests, are untrustworthy, usually suggesting "discriminant validity" even for nearly collinear LV's.

Discriminant validity in the (original) Campbell and Fiske (1959) sense (low correlations with conceptually distinct LV's) could be viewed in terms of the amount of unique err-free variance in correlated LV's after scaling, and the potential for instability of structural coefficient estimates. Stated differently, do AVE's and the covariance between two LV's combine to reduce the unique error-free variance of either (or both) (after scaling) to less than 50%, and thus increase the instability potential of these LV's structural coefficients? In two of the three examples above, AVE's and correlations combined to reduce unique error-free variance after scaling to questionable levels for "reproducibility" in Campbell and Fiske's (1959) terms.

For emphasis,

o with or without obvious discriminant validity problems in the Fornell and Larker sense (i.e., failure(s) of Fornell and Larker's AVE's-versus-squared-correlation (discriminant validity) test), lower AVE's can produce discriminant validity/reproducibility problems (i.e., AVE is insufficient to avoid an increased potential for structural coefficient instability) (see Example 1 above).

o Lower AVE LV's should be investigated for the possibility that their unique error-free variances are less than 50%. There probably can be no firm rule, but Fornell and Larker's suggestion that AVE should be above 0.50 may be insufficient with independent variable correlations above 0.30. Experience suggests that unique error-free variances should be investigated in all correlations larger than 0.7 (again see Example 1 above), especially if an AVE of the LV's involved is less than 0.60.

o Discriminant validity problems should be addressed by raising AVE, not by scaling.

o And finally, any unremedied low unique error-free variance problem should be discussed in the Limitations section of the study's paper, and any discussion of the implications of a significant unremedied low unique error-free variance LV should be prefaced with a caveat that these results are based on more than 50% unique error variance, and thus may be an artifact of the study.

References

Anderson, James C. and David W. Gerbing (1988), "Structural Equation Modeling in Practice: A Review and Recommended Two-Step Approach," *Psychological Bulletin*, 103 (May), 411-23.

Bagozzi, Richard P. and Lynn W. Phillips (1982), "Representing and Testing Organizational Theories: A Holistic Construal," *Administrative Science Quarterly*, 27 (September), 459-489.

Campbell, D. T. and D. W. Fiske (1959), Convergent and Discriminant Validation by the Multitrait-Multimethod Matrix," Psychological Bulletin, 56, 81-105.

Efron, B. (1981), Nonparametric Estimates of Standard Error: The Jackknife, the Bootstrap, and other Resampling Methods. *Biometrika*, 68, 589-599.

Fornell, Claes and David F. Larker (1981), "Evaluating Structural Equation Models with Unobservable Variables and Measurement Error," *Journal of Marketing Research*, 18 (February), 39-50.

Lance, Charles E. (1988), "Residual Centering, Exploratory and Confirmatory Moderator Analysis, and Decomposition of Effects in Path Models Containing Interactions," *Applied Psychological Measurement*, 12 (2) (June), 163-175.

(End)

QUESTIONS of the MOMENT...

"How does one specify and estimate latent variables with only 1 or 2 indicators?"

(The **APA citation** for this paper is Ping, R. A. (2008). "How does one specify and estimate latent variables with only 1 or 2 indicators?" [on-line paper]. http://www.wright.edu/~robert.ping/Under_Det.doc)

One view of a latent variable X, where X is a manifest, observed, continuous, single-indicator, etc. variable, is that X is still a latent variable, but it has only 1 indicator. The loading of x, X's indicator, is 1, and the measurement error of x is 0.

The only difficulty is that it is well known that X seldom has zero measurement error (see Nunnally 1993). As a result, the reliability of X is overstated because the reliability of x is rarely 1, and the structural coefficient in X's association with Y, for example, is typically biased in real-world data. To account for this, one approach would be to relax the assumption of zero measurement error in x. However, X with a single indicator is underdetermined, and a input value for the loading and the measurement error of x must be provided. There has been some confusion over the next steps because the loading and measurement error variance are not independent of each other. The relationship between them involves ρ_x, the assumed reliability of x, and the familiar $e_x = Var_x*(1 - \rho_x)$, where e_x is the measurement error variance of x, and Var_x is the error-attenuated variance of x (e.g., from SAS, SPSS, etc.). In addition, a well-known estimate of ρ_x is the square of the loading of x on X (see for example Bollen 1989). This estimate is exact for standardized Var_x ($Var_x = 1$). Thus, the loading of x and the corresponding measurement error variance of x vary together, and they depend on the choice of the assumed reliability, which usually ranges from 0.7 to 1 (e.g., for an assumed reliability of 0.7 for x, the corresponding loading of x on X is the square root of 0.7, 0.837, and the measurement error variance is computed using $Var_x(1 - 0.7)$ or 0.3 if Var_x is standardized).

To make things simple, consider testing just the "theory-testing extremes" of reliability, $\rho_x = 0.7$ and $\rho_x = 1$, to see if the structural coefficient of the X-Y association becomes non significant (NS) with either of these choices. In particular, to estimate the X-Y structural coefficient at $p_x = 1$, specify X's loading (λ_x) with the square root of $\rho_x = 1$ ($\lambda_x = SQRT(1) = 1$), and specify X's measurement error variance (e_x) with $Var_x*(1 - \rho_x) = 1*(1 - 1) = 0$ ("*" denotes multiplication) in the structural model. Next, repeat this process using $\rho_x = 0.7$ ($\lambda_x = SQRT(0.7)$ and $e_x = Var_x*0.3$). If neither structural coefficient of X-Y is NS at these "extreme" ρ_x's, the safest approach is probably to be conservative and interpret the smaller of the two associations (with the caveat and limitations suggested below).

However, if either of the structural coefficients of X-Y become NS with these "extreme" ρ_x's, there are several possibilities. If a structural coefficient is NS at reliabilities of 0.7 and 1 it probably should be judged to be zero in the population. If it is NS at $\rho_x = 0.7$ and significant at $\rho_x = 1$, the conservative approach would be to judge the NS association(s) to be very likely to be zero in the population. This is because the reliability of X might actually be less than 0.7. If it is NS at 1 and significant at 0.7, the conservative approach again would be to judge the NS association(s) to be somewhat likely to be zero in the population. This is because there is some chance the reliability of X might actually be high.

Thus, if either of the structural coefficients of X-Y is significant for one of the reliability "extremes" and non significant (NS) for the other, that association should probably be judged to be zero in the population. However, depending on the model, this may not be a fatal blow. The lack of significance is likely due to a small standardized structural coefficient for the NS association, and thus this association would not be comparatively "important" to helping explain variance in Y. To practitioners this result could actually be as or more important than a "confirmed" association.

Unfortunately, however, there is more. There is a risk that the reliability of X is less than 0.7 (see below). This may be why theory testers prefer to avoid manifest variables if they can. The possibility that the reliability of X could be less than 0.7 should obviously be stated as a study limitation, and it should be a caveat to any interpretations or implications involving the X-Y association, even if it was significant at both "extremes" of reliability. In addition, because the reliability of X is actually unknown in the study, in the strictest sense this suggests that the X-Y association observed in the study should not be trusted. Thus, the study limitation and caveats that attend manifest variables are (or should be) very serious.

What are the options if the study is complete and X is a focal variable? In general there are several alternatives. These include ignoring X's reliability issue and hoping that reviewers will too (this is not recommended, however, because readers might notice it after publication and dismiss the study), performing a reliability study, and "argumentation." Unfortunately, the term "reliability study" has several meanings. Reliability studies for manifest variables in theoretical model tests *should* involve estimates of intra- and inter-subject rating or measurement of the manifest variables. However, I have not found anything that might be useful in theoretical model tests yet. In the meantime, consider reading the material on "Scenario Analysis" in the *Testing Latent Variable Models Using Survey Data* monograph on this web site. It may be possible to use Scenarios with students to provide multiple inter-subject or inter-subject estimates of X, or a surrogate for X, from which its reliability could be roughly estimated.

A plausible argument might be used to limit the possibilities for the amount of error in X, and provide a rough estimate of its reliability. For example, in several social science literatures the length of the relationship (LENG) is negatively associated with relationship exiting. However, LENG is usually measured in years, which obviously contains measurement error. Nevertheless, the "true" value of LENG for each respondent, informant, or subject (case) is unlikely to be more than about 10 years different from the "observed" or reported LENG in each case. Thus, one more "observation" of LENG could be computed in the data set by adding a (uniform) random number from -10 to 10 to LENG in each case to create LENG_T, an artificial "true" value for LENG. The coefficient alpha of the resulting (artificial) "measure" with the items LENG and

LENG_T *might* be successfully argued to be a plausible estimate of the reliability of LENG. "Might" of course would depend on the reviewers.

As you might suspect the "reliability" depends on several attributes of the distribution of LENG, for example. The coefficient alpha of LENG and LENG_T in a real-world data set of committed relationships (mean = 13.46 years, maximum = 76) was 0.9560. In the same data set the "reliability" of the reported number of employees, EMPL (mean = 7.78, maximum=167), that could be argued to be "off" by 10, 20 or 50 employees, was 0.9612, 0.8693, and 0.5299 respectively. A "better" "reliability" estimate might involve averaging the reliabilities produced by a 100 replications of this procedure.

A slightly different approach might involve estimating a range of values such as those for EMPL, and picking the most conservative, likely, etc. However, this could be labeled "not good science," because "the argument/hypothesis" should always come first in theory testing. Because there are additional difficulties with this range approach, consider resisting it and develop a plausible argument for one "different from" number instead. I would chose 10 for EMPL because experience suggests that in the real world most people know these things and the mean for EMPL was 7.78.

If X has 2 indicators, its specification is slightly less tedious. Nevertheless, X with 2 indicators is still underdetermined, and input values for 2 of X's five measurement parameters must be provided. Alternatively, x_1 and x_2 could be averaged to create a single indicator for X, avg(x_1,x_2) (see Baggozzi and Heatherton 1994). This single indicator would have a loading that is equal to the square root of the reliability of avg(x_1,x_2) (see Bollen 1989). Its measurement error variance is the familiar $e_{avg(x_1,x_2)}$ = $Var_{avg(x_1,x_2)}$*(1 - $\rho_{avg(x_1,x_2)}$), where $e_{avg(x_1,x_2)}$ is the measurement error variance of avg(x_1,x_2), and $Var_{avg(x_1,x_2)}$ is the error-attenuated variance of avg(x_1,x_2) (e.g., from SAS, SPSS, etc.).

It is easy to show that the latent variable reliability of avg(x_1,x_2) is the reliability of the 2 indicators x_1 and x_2 (see Werts, Linn and Jöreskog 1974 for a proposed formula for latent variable reliability). In addition, Anderson and Gerbing (1988) noted that coefficient alpha is practically equivalent to Latent Variable reliability. Thus, to specify X with 2 indicators, the reliability of x_1 and x_2 ($\rho_{avg(x_1,x_2)}$) should be determined using coefficient alpha, x_1 and x_2 should be averaged to produce avg(x_1,x_2), and the error attenuated variance of avg(x_1,x_2) ($Var_{avg(x_1,x_2)}$) should be determined using SAS, SPSS, etc. Next, X should be specified with the single indicator avg(x_1,x_2), the loading of which is the square root of $\rho_{avg(x_1,x_2)}$, and the measurement error variance of which is $Var_{avg(x_1,x_2)}$*(1 - $\rho_{avg(x_1,x_2)}$).

Because coefficient alpha is not identical to latent variable reliability, any t-values involving X (e.g., in an X-Y association) probably should have a significance cut off higher than |t| = 2 (e.g., in the X-Y association with X having 2 indicators, significance might be suggested by |t| greater than or equal to 2.1).

REFERENCES

Anderson, J. C. and Gerbing, D. W. (1988). Structural equation modeling in practice: a review and recommended two-step approach. *Psychological Bulletin*, 103, 411-23.

Bagozzi, R. P. and Heatherton, T. F. (1994). A general approach to representing multifaceted personality constructs: application to self esteem. *Structural Equation Modeling*, 1, 35-67.

Bollen, Kenneth A. (1989), *Structural Equations with Latent Variables*, New York: Wiley.

Nunnally, Jum C. (1993), *Psychometric Theory*, 3rd Edition, New York, NY: McGraw-Hill.

Werts, C.E., R.L. Linn and K.G. Jöreskog (1974), "Intraclass Reliability Estimates: Testing Structural Assumptions," *Educational and Psychological Measurement*, 34, 25-33.

Tmplmvbr.xls

The excel spreadsheet file titled “Tmplmvbr.xls” has been added to the metadata record for “Theoretical Model Testing with Latent Variables” as an additional file.

Please download this additional file to access the original spreadsheet.

EXCEL Template for Computing the (Measurement Error) Adjusted Covariance Matrix for Latent Variable Regression

(The APA citation for this paper is Ping, R. A. (2008). "EXCEL Template for Computing the (Measurement Error) Adjusted Covariance Matrix for Latent Variable Regression." [on-line paper]. http://www.wright.edu/~rping/ItemWeed.doc) .

This EXCEL spreadsheet adjusts a covariance matrix from SAS, SPSS, etc. involving the latent variable Y, a set of up to 5 other latent variables, A through E, and, optionally, all possible interactions and quadratics involving A through E (i.e., AA, BB, AB, CC, AC, BC, DD, AD, BD, CD, EE, AE, BE, CE, DE) for use in error-adjusted-OLS regression ("latent variable" regression). This may seem like a step backward in structural equation analysis, but there are situations involving latent variables where LISREL, EQS, AMOS, etc. are difficult to impossible to use and (error-adjusted) OLS regression is helpful (e.g., model building where OLS regression's forward selection and backward selection are useful, latent variable models with one or more categorical variables, etc.).

The adjustment uses measurement model parameter estimates for the loadings, measurement error variances and variances associated with the latent variables Y, and A through E. The spreadsheet assumes that Y, and A through E are internally consistent (each of their single construct measurement models fit the data), they have mean centered indicators, and that there are no correlated measurement errors involving any of the latent variables A through E.

To use the spreadsheet, a (full) measurement model containing Y and up to five latent variables of interest should be estimated. Next, the bold entries and the italicized entries on the spreadsheet should be deleted to avoid mixing old data with new data, and the result should be zeroes in most of the non- blank areas of the spreadsheet (these values should correct themselves once new data is entered). Then, the covariance matrix to be adjusted should be created using SAS, SPSS, etc. and the variables of interest. Note that this covariance matrix should be created with Y, the dependent/endogenous variable named first. Next, the measurement model loadings, measurement error variances, and variances for Y and the variables of interest should be entered into the appropriate locations on the spreadsheet (i.e., loadings go in the "lambda" lines, measurement error variances go in the "theta" lines, and measurement model variances/covariances for X and Z go in the "Phi" matrix). These entries will all appear in bold font--un-bolded cells are unrelated to entering measurement model parameter estimates). At this point the adjusted covariance matrix will be available beneath the covariance matrix to be adjusted in the middle of the spreadsheet.

Several comments may be of interest. Obviously deleting old data is important in using this spreadsheet. For emphasis, when this spreadsheet (and the others) are visible on a local computer, it can be saved on that computer for later use (i.e., without going back on line). Thus, it is possible to save a copy of the on line version of the EXCEL spreadsheet locally to be used as a "master copy" for modification, subsequent calculations, saving modified copies, etc. The data that appears in the website version of this spreadsheet is also shown in a reordered form in Tables AE1 and AE2 of the monograph *INTERACTIONS AND QUADRATICS IN SURVEY DATA: A SOURCE BOOK FOR THEORETICAL MODEL TESTING (2nd Edition)*, on this web site. Several entries in

Table AE2 are slightly different from the spreadsheet "Adjusted Covariance Matrix..." entries (e.g., Var(SxA) which is Var(AB) in the "Adjusted Covariance Matrix..." of the spreadsheet) for unknown reasons (possibly transcription errors from the spreadsheet to the Table AE2 matrix-- however, the Table AE2 matrix was used to create the latent variable regression results shown in Tables E, G and H, not the spreadsheet).

weeding1.xls

The excel spreadsheet file titled “weeding1.xls” has been added to the metadata record for “Theoretical Model Testing with Latent Variables” as an additional file.

Please download this additional file to access the original spreadsheet.

EXCEL Template for Obtaining Internally Consistent Subsets of Items

This EXCEL template is intended to help delete items from a measure to produce two or more sets of items that "fit the data" (are internally consistent).

The APA citation for these instructions is Ping, R.A. (2006). "More about the template for obtaining an internally consistent set of items." [on-line paper]. http://www.wright.edu/~robert.ping/weeding.doc.

New measures will almost never "fit the data" using a single construct measurement model without dropping items to attain model-to-data fit. In addition, most well established measures developed before covariant structure analysis (LISREL, AMOS, etc.) became popular also will not fit the data without item weeding.

It turns out that measures used with covariant structure analysis are limited to about six items (see discussions in Anderson and Gerbing 1984, Gerbing and Anderson 1993, Bagozzi and Heatherton 1994, and Ping 2008). One explanation is that correlated measurement errors, ubiquitous in survey data but customarily not specified in covariant structure analysis, eventually overwhelm model-to-data fit in single-construct and full measurement models as indicators are added to the specification of a construct. And, that usually happens with about 6 items per construct.

There are ways around item weeding, such as various item aggregation techniques (see Bagozzi and Heatherton 1994), but many reviewers in the Social Sciences do not like these approaches. Unfortunately, reviewers also may not like dropping items from measures because of concerns over face- or content validity (how well the items "tap" the conceptual and operational definitions of their target construct). One "compromise" is to show the full measure's items in the paper, and assuming the full measure does not fit a single construct measurement model, show one submeasure that does fit the data and is maximally "equivalent" to the full measure in face or content validity. However, to do that, several submeasures are usually required, and finding even one is frequently a tedious task.

This template will assist in finding at least two subsets of items from the target measure that fit the data in a single construct measurement model of the items. The process is as follows. First, exploratory (common) factor analyze the target measure with its items using Maximum Likelihood estimation and varimax rotation. If the measure is multidimensional, start with the Factor 1 items. The other factors and the full measure can be used later.

Next, estimate a single construct (confirmatory) measurement model using the Factor 1 items (if the measure is unidimensional Factor 1 is the full measure). If the first measurement model fits the data item omission is not required. If this measurement model does not fit the data, find the "First Order Derivatives" in the output. (I will assume LISREL 8, which requires "all" on the OU line to produce First Order Derivatives. As far as I know, most other estimation packages produce statistics equivalent to First Order Derivatives. For example in SIMPLIS “First Order Derivatives” are available by adding the line “LISREL Output: FD.”). Paste the lower triangle of First Order Derivatives for "THETA-EPS" into the template making sure you retain the item names so you can figure out which item to drop (see the example on the template). Then find the largest value in the "Overall Sum" column--it will be the same as the "Max =" value in the lower right corner of the matrix.

Now, reestimate the measurement model with the item having the largest "Overall Sum" omitted (call this Reestimation 1). Record the Chi Square and RMSEA values on the spreadsheet for reference. If they are acceptable, use the items in this measurement model as submeasure 1.

There is no agreement on acceptable single construct measurement model fit. I use either a Chi Square that is slightly nonzero for single construct measurement models (e.g., 1E-07, not 0), or an RMSEA that is .08 or slightly below, but many authors would suggest much stronger fit criteria for single construct measurement models.[1]

If the unomitted items do not fit the data, find the "First Order Derivatives" for "Theta-Eps" in the Reestimation 1 output. Paste these into the second matrix in the template, record the Chi-Square and RMSEA values, and reestimate the single construct measurement model (Reestimation 2).

Repeating this process, eventually Chi Square will become nonzero, and after that RMSEA will decline to 0.08 or less (the recommended minimum for fit in full measurement and structural models--see Brown and Cudeck 1993, Jöreskog 1993). This should happen with about 7 or 8, down to about 5, remaining items. If acceptable fit does not happen by about 4 items, an error has probably been made, usually by omitting the wrong item.

Each subset after Chi Square becomes non zero is a candidate subset for "best," but because items are disappearing with each step, these smaller subsets are usually less face valid, and thus the first acceptable subset is usually the preferred one.

To find another subset of items, repeat the above process using "Modification Indices" for "Theta Epsilon." (The SIMPLIS command line is “LISREL Output: MI.”) The theory behind Modification Indices is different from First Derivatives, and a different subset usually results.

Another subset of items usually can be found using reliability. The reliability of all the Factor 1 items is computed using SAS, SPSS, etc., the item that contributes least to reliability is deleted, and the reliability of the remaining items is computed. This process is continued until deleting any item reduces reliability. The remaining items usually will fit the data in a single construct measurement model.

If the full measure was multidimensional, there may be several more subsets found by repeating the above procedures using the full measure's items instead of the Factor 1 items, then using the reliability procedure just mentioned. Experience suggests these subsets are smaller, but they frequently include items from Factor 2, etc. and thus they may be more face valid. This process can also be used on any Factor 2 items, Factor 3, etc.

There are many more subsets that can be found by omitting the next largest "Overall Sum" item instead of the "Max =" item. Specifically, the second largest item in Reestimation 1 could be omitted in place of the largest. Then, continuing as before omitting the largest "Overall Sum" items, The result is frequently a different subset of items that fits the data. Another subset can usually be found using this "Second Largest" approach using modification indices instead of first derivatives. Others can be found omitting the second largest overall sum item in Reestimation 2, instead of Reestimation 1, etc., with or without deleting the second largest in Reestimation 1. This "Second Largest" strategy can also be used on the full set of items.

Experience suggests that there are about N-things-taken-6-at-a-time combinations of items with real world data that will fit the data, where N is the number of items in the full measure (more, if 5, 4 and 3 item subsets are counted). For example, if the original measure has 8 items, with real world data there are about 8!(8-6)!/6! = 112 6-tem subsets of items that might fit the data. While the above strategies will not find all of them, experience suggests they should identify several two subsets that are usually attractive because they are comparatively large (again however, usually with about 6 items) and they should appear to tap the target construct comparatively well.

The above spreadsheet approaches may not always identify the highest reliability subsets of items, but experience suggests the resulting subsets are usually larger and as, or more, face valid than those produced by other approaches. However, with low reliability measures, even though the "First Derivative" or "Modification Indices" subsets should be only a few points lower in reliability than a subset found by, for example, dropping items that contribute lest to reliability, the higher reliability subset may be preferred to a higher face validity subset.

It may be instructive to (re)submit all the subsets found to an item-judging panel for their selection of the "best" subset for each construct.

Other comments: There are exceptions to several of the assertions made above, but this is probably not the place for an exhaustive exposition on item deletion strategies. For emphasis, the template assumes lower triangular matrices. There is an additional example in Appendix E of the monograph, *Testing Latent Variable Models...*, on the web site.

REFERENCES

Anderson, James C. and David W. Gerbing (1984), "The Effect of Sampling Error on Convergence, Improper Solutions, and Goodness of Fit Indices for Maximum Likelihood Confirmatory Factor Analysis," *Psychometrika*, 49, 155-73.

Bagozzi, Richard P. and Todd F. Heatherton (1994), "A General Approach to Representing Multifaceted Personality Constructs: Application to Self Esteem," *Structural Equation Modeling*, 1 (1), 35-67.

Browne, Michael W. and Robert Cudeck (1993), "Alternative Ways of Assessing Model Fit," in *Testing Structural Equation Models*, K. A. Bollen et al. eds, Newbury Park CA: SAGE Publications.

Gerbing, David W. and James C. Anderson (1993), "Monte Carlo Evaluations of Goodness-of-Fit Indices for Structural Equation Models," in *Testing Structural Equation Models*, K. A. Bollen and J. S. Long, eds., Newbury Park, CA: SAGE Publications.

Jöreskog, Karl G. (1993), "Testing Structural Equation Models," in *Testing Structural Equation Models*, Kenneth A. Bollen and J. Scott Long eds., Newbury Park, CA: SAGE.

_______ (2004), "On Assuring Valid Measures for Theoretical Models Using Survey Data," *Journal of Business Research*, 57 (2), 125-41.

ENDNOTES

[1] In my opinion, some authors go too far in real world data with single construct measurement model fit, resulting in unnecessarily small submeasures. There are several issues here, including model fit versus face or content validity, and experience suggests that with real-world data, "barely fits" in single construct measurement models is almost

always sufficient to attain full measurement model fit. Thus, in real world data, subsets of items that each produce a comparatively small but nonzero Chi Square or an RMSEA that is just below .08 are usually "consistent enough" to later produce a full measurement model that fits the data. I prefer the RMSEA criterion because it seems to produce fewer problems later. Again, however, many authors would not agree with this strategy. Later, if it turns out that the full measurement model does not adequately fit the data, simply estimate the next item weeding single construct measurement model and drop the next largest "Overall Sum" items to improve full measurement model fit.

ON THE MAXIMUM OF ABOUT SIX INDICATORS PER LATENT VARIABLE WITH REAL-WORLD DATA

Robert Ping
Associate Professor of Marketing
College of Business Administration
Wright State University
Dayton, OH 45435
(937) 775-3047   (FAX) -3545
rping@wright.edu

ON THE MAXIMUM OF ABOUT SIX INDICATORS PER LATENT VARIABLE WITH REAL-WORLD DATA

ABSTRACT

Authors have noted that consistent latent variables have a maximum of about six indicators each. This paper discusses this perhaps surprising behavior and its implications, and an explanation is offered. Approaches to utilizing more than about six indicators in latent variables are also discussed, and several novel approaches are proposed. Each of these approaches is explored using real-world data.

Theoretical model tests (hypothesis testing) involving structural equation analysis combine unobserved or latent variables with proposed linkages among these variables (a model) and proposed (observed) measures of these unobserved variables. These model tests usually involve several steps including defining the model constructs, stating the relationships among these constructs, developing appropriate measures of the constructs, gathering data using these measures, validating these measures, and validating the proposed model.

Commenting on step three, developing appropriate measures, authors have noted that latent variables seem to have an upper limit of about six indicators each (Anderson & Gerbing, 1984; Gerbing & Anderson, 1993; Bagozzi & Heatherton, 1994; Ping, 2004). This apparent "maximum" for latent variable itemization has produced an unfortunate result. Cattell (1973) commented that measures used with structural equation analysis tend to be "bloated specific" (operationally narrow) instances of their target construct. Larger well-established measures developed before structural equation analysis became popular have virtually disappeared from published theoretical model tests involving latent variables. When they do appear in published studies involving structural equation analysis, they frequently are "shadows of their former selves" because of extensive item weeding (i.e., the deletion of items from a measure to attain model-to-data fit).

This paper explores the apparent ceiling of about six indicators per latent variable. An explanation for this result in real-world data is proposed, and approaches to avoiding this apparent limit in theoretical model testing are explored using real-world data.

The observed upper limit of about six indicators per latent variable in published model tests is apparently the result of persistent model-to-data fit difficulties (i.e., inconsistency,[1] see Anderson & Gerbing, 1982) with itemizations containing more than about six indicators per latent variable in real-world data. Gerbing and Anderson (1993) commented that "...fit indices indicated less fit as the...number of indicators per factor, increased..." They went on to propose that "Models with few indicators per factor...have fewer df (degrees of freedom), leaving more 'room to maneuver' the parameter estimates so as to minimize the fit function, which in turn is a function of the residuals."

An Additional Explanation

Intuitively, lack of model-to-data fit in a set of items is the result of unrelated items in that set of items--items that do not "cluster" well enough with the other measure items. Mechanically, the input correlation between an unrelated item in a measure and each of the other measure items cannot be satisfactorily accounted for by the model paths connecting them.[2] Gerbing and Anderson's (1993) comment above suggests that "unrelatedness" increases simply by specifying additional indicators.

The Footnote 2 equation for the model-implied covariance of two unidimensional items suggests an alternative explanation for increased "unrelatedness" when an additional indicator is added to a latent variable. Specifying indicators without accounting for correlations among measurement errors in real-world data (e.g., because of common method) may eventually ruin model-to-data fit. In different words, by ignoring the potential for correlated measurement errors in real-world data, and thus not specifying them, the sum of the residuals (i.e., the sum of the differences between the Footnote 2 computed covariances of the items without the correlated error terms, and the input covariances) eventually becomes unacceptably large.

Next, we will discuss several remedies for lack of model-to-data fit, which will subsequently be investigated using real-world data.

Classical Remedies for Lack of Fit

Classical remedies for lack of model-to-data fit include removing items (item weeding), and correlating indicator measurement errors. The pros and cons of each of these remedies are discussed next.

Item Weeding In published theoretical model tests involving structural equation analysis and real-world data, the about-six-indicators limit frequently produces "item weeding," the removal of items from a measure, to attain a set of indicators that fits the data. This approach has the benefit of producing a subset of items that "clusters" together (i.e., their single construct measurement model is consistent; it fits the data).

However, because the items to be deleted are usually unknown beforehand, item weeding usually capitalizes on chance. In addition, the process of weeding is tedious. As we will see, there may also be several subsets of items for a latent variable that will fit the data (i.e., item weeding may be indeterminate). Finally, structural coefficients, standard errors, and thus observed significances and their interpretation, can vary across these weeded itemizations (i.e., the interpretation of structural coefficients in a model involving weeded subsets can be equivocal).

Item weeding to attain model fit in valid and reliable measures has also been criticized because it impairs content or "face" validity[3] (e.g., Cattell, 1973, 1978; see Gerbing, Hamilton & Freeman, 1994). As mentioned earlier, Cattell (1973) remarked that the resulting weeded measures tend to be bloated specific (operationally narrow) instances of their target construct.

Correlated Measurement Errors It is well known that correlating measurement errors can improve model-to-data fit. This result becomes apparent by examining the Footnote 2 Equation. Including a non-zero correlated measurement error term can improve the model-implied (computed) covariance estimate, and thus it can reduce the corresponding residual. The use of correlated measurement errors presumably to improve fit has been reported (e.g., Bagozzi, 1981a; Byrne & Shavelson, 1986; Bearden & Mason, 1980; Duncan, Haller & Portes, 1971; Reilly, 1982), although this approach has become increasingly rare in recent published model tests. It has the benefit of producing a (sub)set of items that appears to "cluster" together (i.e., their single construct measurement model is consistent; it fits the data). However, as we will see, the indiscriminant use of correlated measurement errors can result in a set of items that appears to be consistent but is actually multidimensional (see Gerbing & Anderson, 1984).

Authors have criticized the use of correlated measurement errors to improve fit (e.g., Bagozzi, 1983; Fornell, 1983; Gerbing & Anderson, 1984) for several reasons. These include that it is a departure from the assumptions underlying classical test theory and factor analysis, and the correlated measurement errors that are specified are typically unhypothesized and thus discovered by capitalizing on chance. In addition, the process of identifying measurement errors that should be correlated is tedious, and, as we will see later, there may be several sets of correlated measurement errors that will produce model-to-data fit (i.e., the results of correlating measurement errors may be indeterminate).

Recent Remedies for Lack of Fit

Comparatively recent remedies for lack of model-to-data fit include using second-order constructs, and aggregating items. These remedies are discussed next.

Second-Order Constructs Gerbing and Anderson (1984) argued that a second-order construct is an alternative to using correlated measurement errors.[4] They suggested that a pair of items with correlated measurement errors could be re-specified as a factor (i.e., as a latent variable, and without using correlated measurement errors), and that a second factor containing the rest of the items, along with the first factor, could be specified as the "indicator" latent variables of a second-order construct. This approach has the benefit of producing a set of items that in their second-order specification fits the data.

However, because the items that should be specified in the first factor are unknown beforehand, the process of identifying these first-factor items could be viewed as capitalizing on chance. In addition, the process of identifying the first-factor items is tedious, and there may be several second-order constructs that will fit the data (i.e., the results of this approach may be indeterminate).

Aggregation Kenny (1979) is apparently credited with an approach that involves summing items in a measure to provide a single indicator of a latent variable. The approach uses reliabilities for loadings and measurement error variances, and variations of this approach have been used in the Social Sciences with structural equation analysis presumably to avoid item weeding. (e.g., Heise & Smith-Lovin, 1981; James, Mulaik & Brett, 1982; Williams & Hazer, 1986).[5]

This full or total aggregation (Bagozzi & Heatherton, 1994) alternative to item weeding has several merits including that it allows the use of older well-established measures having more than six items with structural equation analysis (e.g., Williams & Hazer, 1986).

An assumption in structural equation analysis is that the indicators are continuous. When it is averaged, a summed indicator produces a more-nearly-continuous indicator (e.g., averaged ordinal-scaled indicators then take on ratio-valued numbers) that better approximates this continuous data assumption, and thus an aggregated indicator can reduce the bias that attends the criticized use of structural equation analysis with ordinal (e.g., rating scale) data (e.g., Bollen, 1989; Jöreskog & Sörbom, 1996).

A summed indicator also reduces the size of the input covariance matrix (i.e., the input covariances of a summed indicator replace the input covariances of the several indicators comprising the sum), thus reducing the asymptotic incorrectness of the input covariance matrix for a given sample size. In different words, this helps enable the use of the methodological small samples typical in survey-model tests in the Social Sciences (e.g., 200-300) with larger structural models by improving the ratio of the sample size to the size of the covariance matrix.[6] The use of summed indicators also separates measurement issues from model structure issues in structural equation models. In different words, for an unsaturated structural model, lack of fit with a summed-indicators model unambiguously suggests structural model misspecification, rather than suggesting a combination of measurement model difficulties and structural model misspecification.

However, the indiscriminant use of summed indicators could produce a summed item that is composed of multidimensional items. Summed indicators are also non-traditional in structural equation analysis, and their use could be viewed as not particularly elegant when compared to multiple indicator specification. Further, it is believed that a reliability loading can underestimate the loading of a summed item.

Other Remedies

There are several other remedies for lack of model-to-data fit, including partial aggregation, gauging external consistency only, and using measure validation studies. These remedies are discussed next.

Partial Aggregation Bagozzi and Heatherton (1994) also used partial aggregation--items were grouped into subsets and each subset was summed. This approach avoids the use of reliability loadings used in full aggregation if three or more consistent subsets of items can be found. This approach also has all the benefits and drawbacks of full aggregation. However, because the items that should be aggregated are unknown beforehand, partial aggregation could be viewed as capitalizing on chance. The process of finding consistent subset of items is also tedious, and there may be several aggregations of items that will fit the data (i.e., the results of partial aggregation may be indeterminate).

External Consistency Only An additional alternative to item weeding would be to weed (unidimensional) measures jointly, instead of weeding them singly. Item weeding is typically performed one measure at a time (i.e., using single construct measurement models--see Jöreskog, 1993) to establish the internal consistency of each measure (i.e., each measure fits its single construct measurement model--see Anderson and Gerbing, 1988). Later, the resulting internally consistent (unidimensional) measures are jointly specified in a full measurement model (i.e., a measurement model that contains all the measures) to assess the external consistency of the (unidimensional) measures (i.e., the measures jointly fit a unidimensionally specified full measurement model--again see Anderson and Gerbing 1988). However, it could be argued that the ultimate objective of item weeding is for a full (unidimensionally specified) measurement model to fit the data (to isolate any structural model fit problems to the structural paths among the latent variables). Thus, it should be possible to accomplish full measurement model fit by using measures that are unidimensional in the exploratory common factor sense, omitting the internal consistency evaluation step, and item-weeding using a full measurement model only.

Although this remedy has not been used as far as we know, it should have the benefit of producing measures with fewer items weeded out. However, because the items that should be weeded are typically unknown beforehand, this alternative could be viewed as capitalizing on chance. In addition, the process of weeding is tedious, and there may be several sets of the resulting items that will fit the data (i.e., the results of this weeding may be indeterminate). Further, skipping the internal consistency step violates the current received view in theoretical model testing using survey data: Anderson and Gerbing's (1988) "Two-Step" approach to model respecification in order to attain model-to-data fit (i.e., first verify internal consistency, then verify external consistency).

Measure Validation An approach that might reduce some of the drawback of the above approaches would be to conduct a measure validation study. Ideally, measure validation uses several data sets, one to show measure adequacy (i.e., acceptable psychometrics), and one more to validate (i.e., disprove) the adequacy of the measure.

A measure validation approach might allow the "discovery" of "the" (content valid) weeded subset of items for each measure, "the" (acceptable) second-order construct structure, "the" partial aggregation structure, "the" correlated measurement error structure, or "the" external-consistency-only structure of a measure in study one, and the (dis)confirmation of that structure could be attempted in study two. Thus, this approach might permit the use of weeded subsets, second-order constructs, etc. with less criticism because capitalizing on chance would be removed in the second study.

EXAMPLES

To investigate their efficacy, the above approaches were used with a real-world data set. A mailed-out survey used in a theoretical model test produced more than 200 usable responses. Among the variables in the hypothesized model was the construct N that was measured using a new 18-item measure.[7] While the measure for N was judged to be content or face valid, it was multidimensional (i.e., it had three dimensions using maximum likelihood exploratory common factor analysis). The items in the first factor were subsequently judged to be valid and reliable (the coefficient alpha for Factor 1 was .963), but the single construct measurement model for the Factor 1 items was inconsistent (i.e., it was judged to not fit the data using a single construct measurement model--chi square/df/p-value/RMSEA/GFI/AGFI = 270/35/0.0/.227/.670/.481).[8]

Appendix A provides an example of *item weeding* to produce a subset of consistent items for N. In summary, a total of 20 consistent but different weeded subsets of the items of N were found using a procedure suggested by Ping (1998) (see Ping, 2004). The search for additional weeded subsets was discontinued after it became difficult to determine which weeded subset had the "best" content validity.

Because the weeded items were unknown beforehand, the resulting consistent subsets of items all capitalized on chance. In addition, the process of weeding was very tedious, and because the weeding produced multiple itemizations, the resulting subsets of weeded items were indeterminate.[9] Finally, in a simple structural model of the antecedents of N, one of the structural coefficients, its standard error, and thus its significance, became non significant as alternative weeded itemizations of N were specified. Thus, the interpretation of the structural model involving weeded itemizations of N was equivocal.

Appendix B provides an example of the use of *correlated measurement errors* to produce a set of items for N that fits the data. In summary, two sets of correlated measurement errors were found that resulted in all of the Factor 1 items fitting a single construct measurement model for N. An efficient procedure for finding these correlated measurement errors was discovered, and this procedure was used to find a third set of correlated measurement errors that permitted the full 18 item set to fit a single construct measurement model for N.

Because the measurement errors that were correlated were unknown beforehand, these correlated measurement errors capitalized on chance. Further, the process of identifying measurement errors that should be correlated was tedious. There were also several sets of correlated measurement errors, and thus the resulting sets of correlating measurement errors were indeterminate.

Appendix C probed the use of *second-order constructs* to enable model-to-data fit using the Factor 1 items from the measure for N. In summary, no second order specification of Factor 1 could be found that fit the data without resorting to correlated measurement errors. This suggests that with real-world data a second-order specification for inconsistent items may not always be readily apparent. Specifically, these results suggest that in real-world data logically grouping inconsistent items (e.g., Hunter and Gerbing, 1982; Gerbing, Hamilton and Freeman, 1994) and combining weeded and weeded-out items (e.g., Gerbing & Anderson, 1984) in second-order constructs may not always result in a second-order construct that fits the data.

Appendix D provides examples of the use of *full aggregation*. Using factor scores the full 18-item measure for N was aggregated and used to estimate a structural model containing N.

Aggregation was also accomplished with a single averaged indicator for N composed of its Factor 1 items, then a single averaged indicator for N composed of weeded Factor 1 items. In summary, three interpretationally equivalent (i.e., the directions and significances of their structural coefficients were equivalent) full aggregation approaches were reported in addition to the use of factor scores; averaged indicators with averaged LISREL 8 loadings and measurement errors, with averaged maximum likelihood EFA loadings and measurement errors, and with reliability loadings and measurement errors.

Appendix E presents the results of investigating the other suggestions: Partial Aggregation, gauging External Consistency Only to achieve measurement model fit, and the use of Measure Validation.

Using *Partial Aggregation*, several partitionings of the items were investigated. These included logical groupings of all 18 items in the measure for N (i.e., groupings of items that appeared to tap the same facet of N as Bagozzi and Heatherton, 1994 suggested), creating two summed indicators for N from its Factor 1 items (i.e., an indicator that was the average of the weeded Factor 1 items, and another indicator that was the average of the Factor 1 items that were weeded out), creating 3 summed indicators for N from its Factors 1, 2 and 3 items as three averaged indicators for N, subsets of the 18 items of N using maximum likelihood exploratory common factor analysis (EFA) with forced 7, 6, etc. factor solutions, and subsets of the Factor 1 items of N using EFA with forced 7, 6, etc. factor solutions. However, none of these partial aggregations of the items of N fit the data.

This suggests that the specification of a measure with a large number of items may not always be readily apparent using partial aggregation, even when the items cluster together unidimensionally in an exploratory factor analysis (i.e., an acceptable partial aggregation of the Factor 1 items of N could not be found).

Investigating the omission of the internal consistency verification step and achieving model-to-data fit using *External Consistency Only*, we itemized the 9 latent variables in the study with their Factor 1 items. Then we estimated a full measurement model containing all the model latent variables with each set of items specified unidimensionally (each item was specified with only one underlying latent variable). This full measurement model was judged to fit the data without deleting any additional items to attain model-to-data fit.

This suggests that in real-world data omitting the internal consistency verification step for unidimensional items in the maximum likelihood exploratory factor analysis sense may produce a full unidimensionally specified measurement model that fits the data, thus separating measurement from structure as Anderson and Gerbing (1988) and others have stressed.

In order investigate the use of a *Measure Validation* study to avoid some of the above criticisms of item weeding, correlated measurement errors, etc., we conducted a Scenario Analysis. A Scenario Analysis is an experiment in which subjects read written scenarios that portray a situation or scenario in which the study constructs are manipulated. Then they are asked to complete a questionnaire containing the study measures. Unfortunately the protocol used for the scenario analysis produced missing treatments. As a result, while the resulting scenario analysis was useful for assessing reliability and facets of validity, its results were not appropriate for finding "the" (content valid) weeded subset of items for N, "the" correlated measurement error structure, etc. in order to permit the use of weeded subsets, second-order constructs, etc. with fewer of the criticisms mentioned earlier.

Discussion

These results suggest that several of the proposed alternatives to item weeding may not always be useful in real-world data. Second-order constructs failed to perform in the example.

The example also suggested that partial aggregation of a multidimensional measure may not always be an alternative to item weeding in real-world data. Similarly, the example use of a measure validation study with Scenario Analysis did not perform as expected.

Of the alternatives to item weeding discussed, only full aggregation, external consistency only, and correlated measurement errors performed in the example.

Because correlated measurement errors are comparatively rare in recent published model tests, it seems almost pointless to discuss them further. In addition, the example illustrated how they are found by chance. Because there were several sets of correlated measurement errors, the results of correlating measurement errors can be indeterminate, and an unexplored issue is the effect of changes in correlated measurement errors on structural coefficients.

The example suggested that full aggregation might be used to specify a multidimensional measure as a 2nd order construct. Nevertheless, item weeding will probably continue as the preferred approach to attaining measurement model-to-data fit in survey data model tests even though its use has been criticized, and as the examples suggested, the results of these tests and their interpretation may change materially when items are omitted simply to attain measurement model to data fit.

However, the example also suggested an improved approach to item weeding: for each weeded measure find several item weedings that fit the data, then re-convene the item-judging panel to determine which set of items that best taps the conceptual definition for the measure's construct. A through weeding would include the results of weeding the full measure (e.g., the 18 item measure for N), along with "jacknifed" weedings of the full measure (i.e., remove the first item, then weed the rest; replace the first item then remove the second item and weed the rest; etc.). It would also include weedings from pairwise combinations of any factors (e.g., for the items of N, F1 and F2, F1 and F3, and F2 and F3), along with their "jacknives," and weedings from F1 and its "jacknives." A through presentation of the results of weeding to an item-judging panel would include the full measure and the Factor one items, along with the weeded subsets.

If the full measure or its F1 items are judged to be more content valid than any of the weeded submeasures, External Validity Only and full aggregation could be used for that (sub)measure[10] (i.e., other weeded measures might be combined with Externally Valid Only measures or fully aggregated measures).

Several comments may be of interest. Appendix A illustrated the alternative explanation proposed earlier for the apparent ceiling of about six internally consistent indicators: item weeding reduced the number of unspecified but significant measurement error intercorrelations that contributed to the residuals in a single construct measurement model (specified without correlated measurement errors). Specifically, before weeding there were 25 significant modification indices for the correlations between the measurement errors in the Factor 1 items (not reported), and the sum of these modification indices without regard to sign was 474. As each item was weeded (removed), the number of these significant modification indices declined, and so did their sum without regard to sign. Perhaps surprisingly, the resulting consistent weeded subset, Subset 2, had three significant modification indices for the correlations between the remaining measurement errors.

REFERENCES

Anderson, James C. and David W. Gerbing (1982), "Some Methods for Respecifying Measurement Models to Obtain Unidimensional Construct Measurement," *Journal of Marketing Research*, 19 (November), 453-60.

Anderson, James C. and David W. Gerbing (1984), "The Effect of Sampling Error on Convergence, Improper Solutions, and Goodness of Fit Indices for Maximum Likelihood Confirmatory Factor Analysis," *Psychometrika*, 49, 155-73.

Anderson, James C. and David W. Gerbing (1988), "Structural Equation Modeling in Practice: A Review and Recommended Two-Step Approach," *Psychological Bulletin*, 103 (May), 411-23.

Bagozzi, Richard P. (1981a), "Attitudes, Intentions, and Behavior: A Test of Some Key Hypotheses," *Journal of Personality and Social Psychology*, 41 (4), 607-27.

Bagozzi, Richard P. (1981b), "An Examination of the Validity of Two Models of Attitude," *Multivariate Behavioral Research*, 16 (July), 323-59.

Bagozzi, Richard P. (1983), "Issues in the Application of Covariant Structure Analysis: A Further Comment," *Journal of Consumer Research*, 9 (March), 449-50.

Bagozzi, Richard P. and Todd F. Heatherton (1994), "A General Approach to Representing Multifaceted Personality Constructs: Application to Self Esteem," *Structural Equation Modeling*, 1 (1), 35-67.

Bearden, William O. and J. Barry Mason (1980), "Determinants of Physician and Pharmacist Support of Generic Drugs," *Journal of Consumer Research*, 7 (September), 121-30

Bollen, Kenneth A. (1989), *Structural Equations with Latent Variables*, New York: Wiley.

Browne, Michael W. and Robert Cudeck (1993), "Alternative Ways of Assessing Model Fit," in *Testing Structural Equation Models*, K. A. Bollen et al. eds, Newbury Park CA: Sage.

Byrne, B. M. and R. J. Shavelson (1986), "Adolescent Self-Concept: Testing the Assumption of Equivalent Structure Across Gender," *American Educational Research Journal*, 12, 365-85.

Cattell, R. B. (1973), *Personality and Mood by Questionnaire*, San Francisco: Jossey-Bass.

Cattell, R. B. (1978), *The Scientific use of Factor Analysis in Behavioral and Life Sciences*, New Your: Plenum.

Duncan, O. D., A. O. Haller and A. Portes (1971), "Peer Influences on Aspirations: A Reinterpretation," in *Causal Models in the Social Sciences*, H. M. Blalock, Jr. ed., Chicago: Aldane.

Dwyer, F. Robert and Sejo Oh (1987), "Output Sector Munificence Effects on the Internal Political Economy of Marketing Channels," *Journal of Marketing Research*, 24 (November), 347-358.

Fornell, Claes (1983), "Issues in the Application of Covariant Structure Analysis: A Comment," *Journal of Consumer Research*, 9 (March), 443-47.

Fornell, Claes and David F. Larker (1981), "Evaluating Structural Equation Models with Unobservable Variables and Measurement Error," *Journal of Marketing Research*, 18 (February), 39-50.

Gerbing, David W. and James C. Anderson (1984), "On the Meaning of Within-Factor Correlated Measurement Errors," *Journal of Consumer Research*, 11 (June), 572-80.

Gerbing, David W. and James C. Anderson (1993), "Monte Carlo Evaluations of Goodness-of-Fit Indices for Structural Equation Models," in *Testing Structural Equation Models*, K. A. Bollen and J. S. Long, eds., Newbury Park, CA: SAGE Publications.

Gerbing, David W., Janet G. Hamilton and Elizabeth B. Freeman (1994), "A Large-scale Second-order Structural Equation Model of the Influence of Management Participation on Organizational Planning Benefits," *Journal of Management*, 20, 859-85.

Heise, D. R. and L. Smith-Lovin (1981), "Impressions of Goodness, Powerfulness, and Liveliness from Discerned Social Events," *Social Psychology Quarterly*, 44, 93-106.

Hunter, John Edward and David W. Gerbing (1982), "Unidimensional Measurement, Second-Order Factor Analysis and Causal Models," in *Research in Organizational Behavior*, Vol. IV, Barry M. Staw and L. L. Cummings eds., Greenwich CT: JAI Press, 267-320.

James, J. R., S. S. Mulaik, and J. M. Brett (1982), *Causal Analysis*, Beverly Hills: SAGE.

Jöreskog, Karl G. (1970), "A General Method for Analysis of Covariance Structures," *Biometrika*, 57, 239-251.

Jöreskog, Karl G. (1993), "Testing Structural Equation Models," in *Testing Structural Equation Models,* Kenneth A. Bollen and J. Scott Long eds., Newbury Park, CA: Sage.

Jöreskog, Karl G. and Dag Sörbom (1996), *LISREL 8 User's Reference Guide*, Chicago: Scientific Software International, Inc.

Kenny, David (1979), Correlation and Causality, New York: Wiley.

Ping, R. A. (1998), "Some Suggestions for Validating Measures Involving Unobserved Variables and Survey Data," *1998 Winter American Marketing Association Educators' Conference Proceedings*, Chicago: American Marketing Association.

Ping, R. A. (2004), "On Assuring Valid Measures for Theoretical Models Using Survey Data," *Journal of Business Research*, 57 (2), 125-41.

Reilly, Michael D. (1982), "Working Wives and Convenience Consumption," *Journal of Consumer Research*, 8 (March), 407-18.

Rindskopf, David and Tedd Rose (1988), "Some Theory and Applications of Confirmatory Second-order Factor Analysis," *Multivariate Behavioral Research*, 23 (January), 51-67.

Sörbom, D. (1975), "Detection of Correlated Errors in Longitudinal Data," *British Journal of Mathematical and Statistical Psychology*, 28, 138-51.

Steiger, J.H. (1990), "Structural Model Evaluation and Modification: An Interval Estimation Approach," *Multivariate Behavioral Research*, 25, 173-180.

Werts, C. E., R. L. Linn and K. G. Jöreskog (1974), "Intraclass Reliability Estimates: Testing Structural Assumptions," *Educational and Psychological Measurement*, 34, 25-33.

Williams, Larry J. and John T. Hazer (1986), "Antecedents and Consequences of Satisfaction and Commitment in Turnover Models: A Reanalysis Using Latent Variable Structural Equation Methods," *Journal of Applied Psychology,* 71 (May), 219-231.

Wright, Sewell (1934), "The Method of Path Coefficients, *Annals of Mathematical Statistics*, 5, 161-215.

APPENDICES

APPENDIX A--Item Weeding

To investigate item weeding, the full 18-item measure for N was subjected to a procedure for item weeding suggested by Ping (1998) (see Ping, 2004). This procedure uses partial derivatives of the likelihood function with respect to the measurement error terms (Modification Indices for Theta Epsilon in LISREL, LMTEST in EQS). Specifically, a single construct measurement model for the full 18 item (multidimensional) measure of N was specified unidimensionally and with the correlations among the measurement errors fixed at zero. This produced a matrix of modification indices for the fixed correlated measurement errors which was then examined, and the item in that matrix with the largest summed modification index without regard to sign (i.e., the sum of the item's column of modification indices without regard to sign) was deleted. Next, the single construct measurement model without this item was re-estimated, and the item with the largest summed modification index without regard to sign in the resulting modification indices for the correlations among the measurement errors was deleted. This process was repeated, deleting an item at each step, until a subset of the 18 items was found that fit the data.

The resulting 5-item subset (Subset 1--containing the items n_1, n_{12}, n_{14}, n_{15} and n_{18}) was consistent (it fit the data--chi square/df/p-value/RMSEA/GFI/AGFI = 2.84/5/.723/0/.991/.974) (see Footnote 8 for a discussion of model fit), and it contained items from Factor 1 (n_{12}, n_{14}, n_{15} and n_{18}) and an item from Factor 3. There was another consistent 5-item subset (Subset 1a-- n_4, n_{12}, n_{14}, n_{15} and n_{18}--chi square/df/RMSEA/GFI/AGFI = 9.09/5/.105/.079/.973/.919), that contained the Factor 1 items and a Factor 2 item. However, we could not find a consistent subset with items from all three factors. Nevertheless, it was possible to find consistent subsets with items from several factors.

Using this derivative procedure on the 10 items of Factor 1, a different consistent subset obtained (Subset 2--n_9, n_{13}, n_{15}, n_{16}, n_{17} and n_{18}--chi square/df/p-value/RMSEA/GFI/AGFI = 15/9/.086/.072/.961/.910). This 6-item subset was judged to be slightly more content or face valid than Subsets 1 or 1a, and its items clustered together using maximum likelihood exploratory common factor analysis slightly better than Subset 1 (the percent of the variance explained for Subset 2 was 75.7%, versus 59.8% for Subset 1 and 66.8% for Subset 1a).

Several more consistent subsets were then obtained. Obviously any subset of Subset 2 would fit the data, and there were 41 of these (= the total number of combinations of 6 things taken 5, 4 then 3 at a time). In addition, arbitrarily omitting an item from the Factor 1 set of 10 items produced two more consistent subsets, Subset 3 (n_9, n_{10}, n_{11} and n_{17}--chi square/df/p-value/RMSEA/GFI/AGFI = 3.34/2/.187/.072/.987/.939), and Subset 4 (n_{11}, n_{13}, n_{16}, n_{17} and n_{18}--chi square/df/p-value/RMSEA/GFI/AGFI = 2.64/5/.754/0.0/.992/.976). These subsets clustered together about as well as Subset 2 (71.6% and 75.1% explained variance respectively) and judging which had the "best" content validity became impossible without resorting to an item-judging panel. However, we judged each of the weeded measures to be less content valid than either the original 18-item measure, or the 10 item Factor 1 measure.

We then discontinued the search.[11] In summary, we identified 20 consistent subsets of the 18-item measure for N using the derivative procedure.

Finally, we gauged the sensitivity of structural coefficients to changing the itemization of N. In a simple saturated structural model with N as the single endogenous variable, the structural coefficients and standard errors were judged to vary unpredictably across these different itemizations of N. For example, the t-value for one of the 4 significant structural coefficients changed from t = 2.66 to t = 0.91 by changing the itemization of N from weeded Subset 1a to Subset 4.[12]

APPENDIX B--Correlated Measurement Errors

In order to investigate correlated measurement errors, the full 18-item measure for N was subjected to a procedure involving modification indices. A single construct measurement model for the full (multidimensional) measure of N was specified unidimensionally and with the correlations among the measurement errors fixed at zero to produce a matrix of modification indices for the fixed correlated measurement errors. Then, the measurement error correlation corresponding to the largest of these modification indices was freed (i.e., the corresponding measurement errors were allowed to correlate) (a modification index of 3.8 is significant at p = .05 with 1 degree of freedom, see the second part of Footnote 14). Next, the single construct measurement model was estimated with this measurement error correlation freed, and the largest of the resulting modification indices for the remaining fixed correlations among the measurement errors was found and freed. This process was repeated, freeing a measurement error correlation at each step, a total 90 times before we decided to abandon this "forward selection" process of identifying correlated measurement errors.[13]

However, we used the above "forward selection" approach on a smaller subset of items, the Factor 1 items (10 items-- n_9 through n_{18}), until the set of Factor 1 items was judged to fit the data. The Factor 1 items with correlated measurement errors was judged to be consistent (i.e., it fit the data--chi square/df/p-value/RMSEA/GFI/AGFI = 34/20/.022/.074/.949/.861) (see Footnote 8 for comments on model fit). The procedure required 27 estimations and produced 15 significant correlated measurement errors (9:10,11; 10:12,13,16,17; 11:--; 12:13,17; 13:14,15,16;

14:17,18; 15:16,18; 16:--; 17:--; where for example 9:10,11 denotes the correlations between ε_9 and ε_{10}, and ε_9 and ε_{11}, where ε denotes a measurement error term, and 11:-- for example indicates that ε_{11} was not correlated with its higher-ordinality measurement errors, ε_{12} through ε_{18}).

To find another set of correlated measurement errors for the Factor 1 items, we specified a single construct measurement model for the Factor 1 items with all the measurement error correlations fixed at zero, except for the ε_9 correlations which were freed.[14] Estimating this model we recorded the significant measurement error correlations between ε_9 and ε_{10} through ε_{18}. Next we re-fixed the measurement error correlations with ε_9 to zero, and freed the ε_{10} measurement error correlations with its higher-ordinality measurement errors, ε_{11} through ε_{18} (i.e., all measurement error correlations were fixed at zero except for those between ε_{10} and ε_{11}, ε_{10} and ε_{12}, ... , and ε_{10} and ε_{18}). After estimating this model, we recorded the significant measurement error correlations between ε_{10} and ε_{11} through ε_{18}. Repeating this process of re-fixing the previously freed measurement error correlations, and freeing and estimating the higher-ordinality measurement error correlations for ε_{11}, then ε_{12}, ... , then ε_{18} (e.g., the ε_{11} correlations with ε_{12} through ε_{18}, the ε_{12} with ε_{13} through ε_{18}, etc.), the result was a set of significant measurement error correlations for ε_9 through ε_{18}.

Next, we re-specified the single construct measurement model for Factor 1 with the all measurement error correlations again fixed at zero. Then, we freed the significant modification indices just recorded for ε_9, ε_{10}, etc. (i.e., based on their recorded modification indices, the significant correlations for ε_9 were freed, the significant correlations for ε_{10} were freed, etc.). This single construct measurement model was estimated (chi square/df/p-value/RMSEA/GFI/AGFI = 31/14/.004/.098/.956/.828) and the nonsignificant measurement error correlations were trimmed (i.e., fixed at zero--for example 10:11,15 were trimmed because they were nonsignificant when estimated in the presence of the other specified measurement error correlations). This trimmed model was then estimated, and because it did not yet fit the data (chi square/df/p-value/RMSEA/GFI/AGFI = 53/24/.0004/.097/.931/.841) the modification indices (MI's) for the remaining non-freed measurement error correlations were examined to find the largest significant MI, $MI_{9,12}$ (= the modification index for the ε_9-ε_{12} correlation = 10.30). The ε_9-ε_{12} correlation was then freed and the resulting single construct measurement model was estimated. This measurement model was judged to fit the data (chi square/df/p-value/RMSEA/GFI/AGFI = 41/21/.009/.079/.945/.870).

Several comments may be of interest. There were two sets of correlated measurement errors that would permit Factor 1 items to fit the data in a single construct measurement model. Stated differently, there was more than one set of correlated measurement errors that would make the Factor 1 items consistent in a single construct measurement model. While the correlations from the second or "column-wise" selection approach were more parsimonious (i.e., there were fewer of them) and they were found using half as many estimations (12--1 for each item, one more to trim the nonsignificant intercorrelations, plus one to add the additional intercorrelation $MI_{9,12}$--versus 27 for forward selection), their comparative statistics are trivially different (AIC/CAIC/EVCI for the column-wise selection = 105/229/.815 versus 104/240/.803 for forward selection).

Since the second or column-wise selection approach required considerably fewer estimations, we tried it on the full set of 18 items. Obtaining a set of measurement error correlations that were judged to make the full 18-item measure consistent required 19 estimations, one for each item, one to trim the resulting nonsignificant correlations, and 9 more estimations to add enough additional correlations to obtain consistency (chi square/df/RMSEA/GFI/AGFI = 122/69/.00007/.077/.906/.768).

Thus, the original multidimensional set of 18 items could be specified so that it fit the data in a single construct measurement model using correlated measurement errors. Stated differently, correlated measurement errors masked a multidimensional measure.

Finally, the trimming step in the column-wise selection approach (i.e., to remove nonsignificant correlated measurement errors) suggests that some measurement error correlations were collinear. Stated differently, freeing a correlation affected the significance or lack thereof in other correlations. This may explain the apparent indeterminancy in measurement error correlations. Specifically, the starting point (i.e., the measurement errors that were initially allowed to correlate) determined the remaining significant measurement error correlations.

APPENDIX C--Second-Order Constructs

To investigate the use of second-order constructs we tried several approaches. The initial objective was to find a second-order construct for the Factor 1 items that would fit its single construct measurement model. To this end authors have suggested grouping items into subsets using their content or face validity (i.e., grouping items that seem to be related based on their wording--see Hunter and Gerbing, 1982; Gerbing, Hamilton and Freeman, 1994). For example, we grouped the Factor 1 items into two subsets based on wording (i.e., the subset 1 items were used to indicate latent variable 1, the subset 2 items were used to indicate latent variable 2, and latent variables 1 and 2 were specified as the "indicators" of the now second-order Factor 1), then three subsets. However, we were unable to find a grouping of the Factor 1 items based on item wording that fit their single construct measurement model (i.e., the measurement model containing only the second-order construct) without resorting to correlated measurement errors.

Authors have also suggested that weeded-out items might be specified as a second "indicator" factor in a two-factor second-order construct (i.e., a second-order construct with the weeded items as one "indicator" latent variable, and the items that were not weeded out as the other indicator latent variable--see Gerbing and Anderson, 1984). To this end we specified a second-order construct with the weeded-out items as one "indicator" latent variable, and the surviving items as another "indicator" latent variable. Again, we were unable to find a second-order construct that would fit their single construct measurement model without resorting to correlated measurement errors.

We also tried a second-order construct with the two factors that resulted from a forced two-factor solution in maximum likelihood exploratory common factor analysis of the Factor 1 items, then a forced three-factor solution. These second-order constructs also would not fit their single construct measurement model without resorting to correlated measurement errors.

Finally, we tried specifying a second-order construct with three consistent subsets of the Factor 1 items (i.e., two consistent four-item subsets, and one three item subset that fit the data exactly). This second-order construct also did not fit the data without resorting to correlated measurement errors.

Several comments may be of interest. These results suggest that with real-world data the use of a second-order construct may not always easily improve model-to-data fit.

APPENDIX D--Full Aggregation

In full aggregation, a set of items is summed to form a single indicator. Because the resulting latent variable is underdetermined with only one indicator, it requires that two of its three estimated parameters, its loading, its measurement error variance or its latent variable's variance, be fixed for identification.

It is easy to show that the loading of a summed indicator is the sum of its individual indicator loadings, and that its measurement error variance is the sum of the individual indicator measurement error variances.[15]

A reliability loading and the well-known measurement error estimate Variance∗(1-reliability) has been used in the social sciences. It is easy to show that these estimates for an aggregated indicator are exact when the variance of its latent variable is 1 and latent variable reliability is available (see Appendix F).

Ping (2004) suggested using maximum likelihood exploratory factor analysis (EFA) loadings and reliability measurement error variances for a fully aggregated indicator. There is an additional estimate of the measurement error variance available using EFA results (see Equation F6 in Appendix F).

An additional approach would be to replace indicators with their fully aggregated EFA factor scores.

Each of these aggregation alternatives will be explored next.

While it is obviously possible to aggregate a multidimensional measure, none of the above estimates for the resulting single indicator's loading and measurement error variance would be appropriate (because each requires or assumes unidimensionality), with the exception of factor scores. Pursuing that option we produced factor scores using maximum likelihood exploratory factor analysis and the full 18 item measure of N. Specifically, maximum likelihood EFA of N produced three factors, and the factor score for each of these factors was summed then averaged. Next, a full structural model (i.e., with all the latent variables and N as the single endogenous variable) were estimated with N specified using this single aggregated factor score indicator (with a loading of 1 and a measurement error of 0). The resulting structural model was judged to fit the data (chi square/df/RMSEA/GFI/AGFI = 2349/1449/0/.066/.650/.614) (see Footnote 8 for comments about assessing model-to-data fit).

Next, we estimated a series of structural models involving N and the other model latent variables with a full aggregation (average) of the F1 items, or a full aggregation (average) of a weeded subset of the F1 items, Subset 2 from Appendix A (items n_9, n_{13}, n_{15}, n_{16}, n_{17} and n_{18}). The resulting single indicator used the estimates for loadings and measurement error variances mentioned above, averaged LISREL 8 loadings and measurement errors, averaged maximum likelihood EFA loadings and measurement errors, reliability loadings and measurement errors (see Appendix F), and factor scores.

The structural coefficients on the paths to N from the other 8 latent variables were compared to the structural coefficients produced by the equivalent structural model (i.e., containing N and the other 8 latent variables) that used either the 10 items in F1 or Subset 2, with N specified with multiple indicators that were the individual items of F1 or Subset 2.

The Subset 2 full measurement model fit the data (chi square/df/RMSEA/GFI/AGFI = 2789/1733/0/.065/.630/.596), as did its (saturated) structural model (chi square/df/RMSEA/GFI/ AGFI = 2789/1733/0/.065/ .630/.596). The structural coefficients that resulted were used as a basis for assessing the efficacy of the various alternative loadings and measurement error variances mentioned above.

In summary, the factor score indicator produced by a maximum likelihood EFA with just the Subset 2 items (i.e., with no other items present in the EFA) was judged to produce the smallest differences between the corresponding t-values of the Sunset 2 (baseline) structural model and the factor score indicator structural model. The root mean square (RMS) (pairwise) difference of t-values across the 8 structural coefficients on the paths to N was .002 and the average difference without regard to sign (MAD) was .005 (with a range of between .000 for the nonsignificant structural coefficients to .012 for the significant structural coefficients) (structural coefficient RMS = .011, MAD = .024, range = [.004, .066]).

The smallest structural coefficient differences were produced by the averaged LISREL 8 loadings and measurement errors--the RMS difference of structural coefficients across the 8 structural coefficients on the paths to N was .007 and the MAD was .015 (the range was .001 to .041) (t-value RMS = .065, MAD = .149, range = [.001, .353]).

The t-value and structural coefficient differences for the other aggregation approaches were nearly identical to those produced by the averaged LISREL 8 loadings and measurement errors. For example, the average of the maximum likelihood EFA loadings and an Equation F6 measurement error variance produced a t-value RMS, MAD and range of .062, .149 and [.001, .353], respectively. Its structural coefficient RMS, MAD and range were .005, .015 and [.000, .024], respectively. The reliability loading and measurement error was similar (t-value RMS = .065, MAD = .149, range = [.001, .353]) (structural coefficient RMS = .006, MAD = .015, range = [.001, .031]).

These results were then used to predict the ranking of the performance of N specified using the fully aggregated F1 items and the above approaches. As a baseline structural model the External Consistency Only (see Appendix E) (full) measurement model for N using the F1 items was re-specified as a structural model. An External Consistency Only full measurement model uses unidimensional sets of indicators that are not necessarily internally consistent (i.e., their single construct confirmatory measurement model may not fit the data). In the present case, each latent variable in the External Consistency Only full measurement model had a unidimensional itemization (in a maximum likelihood EFA sense), but none of these itemizations was consistent (in the confirmatory factor analysis sense). However, the resulting External Consistency Only full measurement model measurement model fit the data (chi square/df/RMSEA/GFI/AGFI = 3480/1979/0/.073/.590/.556) (see Footnote 8 for comments about assessing model fit), as did the corresponding structural model (chi square/df/RMSEA/GFI/AGFI = 3480/1979/0/.073/.590/.556).

As with the weeded Subset 2 of the measure for N, the factor score indicator was judged to have produced the smallest t-value differences (t-value RMS = .032, MAD = .070, range = [.001, .167]). However, it also produced the smallest structural coefficient differences (structural coefficient RMS = .013, MAD = .022, range = [.000, .085]).

The other approaches investigated produced nearly identical results. For example, averaged LISREL 8 loadings and measurement errors produced t-value RMS, MAD and range of .048, .124, [.066, .236], respectively (structural coefficient RMS = .050, MAD = .093, range = [.005, .300]). Averaged maximum likelihood EFA loadings and measurement errors were similar (t-value RMS = .048, MAD = .122, range = [.066, .236]) (structural coefficient RMS = .050, MAD = .092, range = [.004, .300]), as were reliability loadings and measurement errors (see Appendix F) (t-value RMS = .047, MAD = .120, range = [.066, .236]) (structural coefficient RMS = .050, MAD = .093, range = [.000, .300]).

The details of these estimations were as follows. For the factor score approach, a maximum likelihood EFA of the 10 F1 items (i.e., with all other items absent) was performed, and the resulting factor scores were added to the data set (i.e., the resulting factor scores were saved and given the variable name FS). Next, N was specified using the single indicator FS with a fixed loading of 1 and a fixed measurement error variance of 0 (i.e., the variance of N, PHI_N, was free). The resulting structural model was judged to fit the data (chi square/df/RMSEA/GFI/AGFI = 2327/1449/0/.065/.652/.616) (see Footnote 8 for comments about assessing model-to-data fit).

In the LISREL 8 loadings and measurement errors approach, a single construct measurement model of the 10 F1 items (i.e., with all other latent variables absent) was estimated with the variance of N free (i.e., one indicator loading was fixed at 1 to provide a metric for N). Although this single construct measurement model did not fit the data (chi square/df/RMSEA/GFI/AGFI = 273/35/0/.219/.687/.509), the F1 items were unidimensional using maximum likelihood EFA, so the resulting loadings were averaged and the resulting measurement error variances were divided by 10^2 (using expectation algebra, the variance of an average of independent measurement errors, a constant times the sum of the measurement errors, is the constant squared times the sum of the variances of the measurement errors). Next, the F1 indicators were averaged, then N was specified using the resulting single indicator with a fixed loading and measurement error variance equal to the averaged LISREL 8 loadings and the "averaged" measurement error variances just described, respectively (i.e., the variance of N, PHI_N was free). The resulting structural model fit the data with the same fit statistics as the factor score model.

For the maximum likelihood EFA loadings and its Equation F6 measurement error variance estimation, a maximum likelihood EFA of the 10 F1 items (i.e., with all other items absent) was performed. The resulting loadings were re-scaled then averaged,[16] and the measurement error variance of this single was calculated using Equation F2 in Appendix F. Next, the F1 items were averaged, and N was specified using the resulting single indicator with a fixed loading and measurement error variance equal to the averaged EFA loadings and the Equation F6 measurement error variance just described, respectively (i.e., the variance of N, PHI_N again was free). The resulting structural model also fit the data with the same fit statistics as the factor score model.

In the reliability loading and measurement error approach, the coefficient alpha reliability (α) and the error-attenuated variance (V) (i.e., from SPSS, SAS, etc.) of the F1 items was determined. The F1 items were again averaged to form a single indicator of N, and this single indicator's the measurement error variance was fixed at the Equation F2 value of = V*(1- α), the loading of the single averaged indicator of N was fixed at the square root of the coefficient alpha reliability, and the variance of N, PHI_N was freed. Again the resulting structural model fit the data with the same fit statistics as the factor score model.

Several comments may be of interest. Specification of the weeded and "weeded-out" items of F1 was accomplished using several aggregation approaches.

The successful estimation of a structural model for a multidimensional N specified with aggregated factor scores, suggests that aggregated factor scores might be an alternative to a 2nd order factor or partial aggregation (see Appendix E) for specifying a multidimensional measure in structural equation analysis. In the present case an 18 item measure which formed three factors using maximum likelihood EFA was fully aggregated using (averaged) maximum likelihood EFA factor scores.

Other full aggregation indicators, loadings and measurement error terms were investigated (e.g., normed indicators) but not reported because they were judged to have performed worse than the reported approaches.

The results from reliability and LISREL 8 loadings were a surprise. Fixing the loading to the square root of coefficient alpha overstates the loading (see Equation F3 in Appendix F) which should not have performed well. Similarly, the LISREL 8 single construct measurement model for the items of F1 did not fit the data, yet the resulting loadings and measurement error variances were useful in this investigation. However, the External Consistency Only measurement model for N with the F1 items did fit the data, and the F1 loadings and measurement error variances were trivially different from those produced by single construct measurement model for the items of F1. This suggests an additional full aggregation approach which is identical to the LISREL 8 approach (a), except that it uses loadings and measurement error variances from an External Consistency Only measurement model for N.

Comparing the structural coefficients and t-values for the 18-item measure for N with those from F1 and the weeded Subset 2 (not reported), structural coefficients and significances changed materially when items were removed from an aggregated indicator for N (2 structural coefficients that were significant using the 18 item measure became nonsignificant when items were dropped, and 1 structural coefficient became significant when items were removed). In different words, this also suggests that changes in content or face validity of a measure (i.e., the items included or excluded in a measure) may change the construct validity of that measure (i.e., the correlations among the other latent variables in the model).

Finally, to investigate the sensitivity of structural coefficients and their significances to small changes in itemization with full aggregation, we specified N with fully aggregated weeded Subsets 1 through 4 and 1a (see Appendix A). The resulting sensitivity to different weeded subsets of items observed were similar to those reported in Appendix A. This suggests not only that item changes may change the study results and their interpretation, but that full aggregation may not mask these changes.

APPENDIX E--Other Approaches

Partial Aggregation To investigate partial aggregation, we grouped the items of N into subsets of items based on similar face or content validity (i.e., we grouped items that appeared to tap the same facet of N together) as Bagozzi and Heatherton (1994) suggested. We then specified N with the summed items (i.e., N was specified with 2 or 3 summed indicators). However, neither of these groupings fit a single construct measurement model for N.

Next we created two indicators for N from the results of weeding. We summed the weeded Factor 1 items, then did the same for the Factor 1 items that were weeded out. Using a fixed reliability measurement error variance for the summed weeded-out items because the latent variable was underdetermined, the measurement model was judged to not fit the data (chi square/df/RMSEA/GFI/AGFI = 16/8/.047/.082/.979/.871) (the model fit is close to being acceptable, but N was not unidimensional--modification indices suggested that the "weeded out" indicator loaded significantly on several latent variables).

Next, we created three indicators for N using the summed items from Factor 1 for one indicator, the summed items from Factor 2 for another, and the summed items from Factor 3 for the last indicator. However, N was again not unidimensional in the full measurement model containing N and the other latent variables, and it did not fit the data.

Finally, we tried obtaining subsets of the items of N using maximum likelihood exploratory common factor with forced 7, 6, etc. factor solutions, and each factor's items were then summed. While the 7, 6, etc. forced factor varimax and oblimin exploratory factorings converged with the 18 item measure of N, and the 6, 4 and 3 factor exploratory factorings also converged with the Factor 1 items of N, the forced 7 and 5 factors with the Factor 1 items did not converge. In addition, none of the successful forced factorings fit the data (the 3-factor solution was not unidimensional in the full measurement model).

These results suggest that partial aggregation may not always allow the specification of a large number of items (i.e., more than about 6). Specifically, in the present case none of the partial aggregating approaches produced a full measurement model that was external consistent.

External Consistency Only To investigate omitting the internal consistency step for unidimensional measures and achieving full measurement model-to-data fit using external consistency only, we itemized each of the 9 latent variables with their Factor 1 items. Each latent variable thus had a unidimensional itemization (in a maximum likelihood exploratory factor analysis sense), but none of these itemizations was consistent (i.e., none fit their single construct measurement model). However, a full measurement model containing the 9 latent variables specified unidimensionally with their respective Factor 1 items (i.e., each item was "pointed to" by only one latent variable) was judged to fit the data (chi square/df/RMSEA/GFI/AGFI = 3480/1979/0/.073/.590/.556) (see Footnote 8 for comments on assessing model-to-data fit).

To probe the limits of this externally consistent measurement model we weeded each of the multidimensional measures until they became unidimensional in a maximum likelihood exploratory factor analysis (EFA) of the un-weeded items. Specifically, each multidimensional measure was specified in a single factor measurement model as though it were unidimensional. For each measure the partial derivative technique used in Appendix A was used to weed the first item from that measure. The un-weeded items in that measure were factored using maximum likelihood EFA to check their factor structure. If the un-weeded items were multidimensional another item was weeded using the partial derivative technique, and the factor structure of the resulting un-weeded items was again checked using maximum likelihood EFA. This process was repeated until the measure was unidimensional using maximum likelihood EFA.

A full measurement model containing these larger but unidimensional measures was also judged to fit the data, but LISREL produced a warning message that the sample size was smaller than the number of parameters to be estimated, and that the parameter estimates were thus unreliable.

Measure Validation Measure validation (i.e., the determination of the adequacy--reliability and validity--of a measure) in survey models can take several approaches. These include a separate large scale study(s) aimed solely at validating the study measures. However, presumably because of budget and time constraints, large scale measure validation studies are sometimes bypassed, and measure adequacy is gauged in a small scale pretest survey(s) (e.g., 100 cases). These small pretest surveys are used to preliminarily assess the measures, and to determine response rates. Obviously, weeded subsets, second-order constructs, etc. may be difficult if not impossible to investigate with the resulting small data sets.

Again perhaps because of budget and time constraints, the final- (model-) test data set is sometimes used for measure validation. In this case the final-test data are used for two separate purposes: to assess the measures, and to validate or test the hypothesized model that uses these measures. In this case weeded subsets, second-order constructs, etc. could obtain, but all the earlier criticisms (capitalizing on chance, etc.) then apply.

To investigate the use of a separate large-scale measure validation study that takes less time and is less expensive than to a mailed-out survey, we conducted a Scenario Analysis. A Scenario Analysis is an experiment in which the subjects (usually students) read written scenarios that portray a situation in which the study constructs are verbally manipulated (i.e., high for one experimental subject--"you are *very* satisfied with..."--low for another--"you are *very dissatisfied* with..."). Then the subjects are asked to complete the study questionnaire which contains the measures to be validated. Compared with other research designs such as cross sectional surveys, the results of Scenario Analyses have been reported to be similar enough that they might be useful in measure development and validation (Ping, 2004).

A Scenario Analysis designed to assess N and the other study measures was conducted using students. An audit of the resulting completed questionnaires suggested, however, that many scenarios were incomplete or were not administered. Specifically, while there were nine variables in the proposed model, eight were exogenous, and thus the scenario required $2^8 = 256$ completed questionnaires to produce one questionnaire for each treatment (i.e., treatment_1 = high exogenous variable_1, high exogenous variable_2, ... , high exogenous variable_8; treatment_2 = high exogenous variable_1, high exogenous variable_2, ... , low exogenous variable_8; treatment_3 = high exogenous variable_1, high exogenous variable_2, ... , low exogenous variable_7, high exogenous variable_8; treatment_4 = high exogenous variable_1, high exogenous variable_2, ... , low exogenous variable_7, low exogenous variable_8; ... ; treatment_{256} = low exogenous variable_1, low exogenous variable_2, ... , low exogenous variable_7, low exogenous variable_8). However, substantially fewer than 256 usable questionnaires were obtained, and re-administering the missing scenarios was judged to be out of the question because it was nearly impossible to determine which scenarios were missing.[17] The effect of these missing treatments was subsequently judged to be unknown.

However, comparing the results of single construct exploratory common factor analysis of each measure, of the 9 measures, the behavior of 7 measures was same between the scenario analysis and the final test: 5 measures were unidimensional in both studies and 2 measures were multidimensional in the two studies, while 2 measures were unidimensional in scenario analysis and multidimensional in the final test, and no measures were multidimensional unidimensional in scenario analysis and unidimensional in the final test (not reported). However, loadings in the unidimensional measures were different between the data sets (i.e., an item that loaded high in the scenario data loaded lower in the final test data, and vice versa). Further, the multidimensional factor structure was not constant across the data sets--the number of factors were usually different between the data sets, but items in the scenario factor 1's were contained in (i.e., a subset of the) final test factor 1's in all but 1 case.

Further, reliabilities were within a few points of each other between the two data sets. While Average Extracted Variances (AVE's) (see Fornell & Larker, 1981) varied more widely (1 to 19 points), when the scenario AVE's were above .5 so were the final test AVE's.

Encouraged by these sanguine results despite the missing treatments, we weeded the measure for N using the scenario data. As discussed in Appendix A, the first weeded subset of the Factor 1 N items in final test data was Subset 2 (n_9, n_{13}, n_{15}, n_{16}, n_{17} and n_{18}). However, n_{16} was the first item weeded out of Factor 1 in the scenario data (not reported). Thus, for the focal construct N, weeded subsets would not (all) be the same across data sets, and finding "the" weeded subset of items for the Factor 1 items of N using this scenario data set was judged unlikely.

Similarly, the first set of correlated measurement errors found for the Factor 1 items of N in the final test data was (9:10,11; 10:12,13,16,17; 11:--; 12:13,17; 13:14,15,16; 14:17,18; 15:16,18; 16:--; 17:--; where for example 9:10,11 indicates the correlations between ε_9 and ε_{10}, and ε_9 and ε_{11}--ε is a measurement error term, and 11:-- for example indicates that ε_{11} was not correlated with its higher-ordinality measurement errors, ε_{12} through ε_{18}). However, 17:18 was the first correlated measurement error identified in the Factor 1 items of N using the scenario data (ε_{17} and ε_{18} were not correlated in the Factor 1 items in the final test data). Thus, the correlated measurement errors for N would not all be the same across data sets, and finding "the" correlated measurement errors for the Factor 1 items of N using this scenario data set was also judged to be unlikely.

We did not investigate 2nd order constructs or partial aggregation because they did not perform in the final test data. Similarly, because the itemizations of Factor 1 in nearly half of the study latent variables were different between the two data sets, weeding using external consistency only was also not investigated in the scenario data.

Since the reliabilities were similar between the two data sets, however, we did investigate full aggregation. However, the correlations were not same between data sets. For example, N had a correlated with S, an important study variable, of .10 in the scenario data, but this correlation was -.54 in full test data. Thus, the proposed structural model was not investigated in the scenario data because structural coefficients are related to partial correlations (and the scenario data was intended for measure validation rather than model validation).

In summary, while this scenario analysis may have been useful for assessing N and the other study measures' reliability and facets of validity (even with missing treatments), its results were not appropriate for finding "the" (i.e., a content valid) weeded subset of items for N, "the" correlated measurement error structure, etc. in order to permit the use of weeded subsets, second-order constructs, etc. with fewer of the criticisms mentioned earlier because capitalizing on chance would be removed in the second study. However, it remains an open question whether or not a "proper" scenario analysis (i.e., one in which all the treatments were administered) would have produced the same conclusion regarding weeded subsets, etc.

Unfortunately, since one was not performed for this analysis (or the original model test, due to budget and time constraints), it is also an open question whether a large scale measure validation study could have been used to find "the" weeded subset of items for N, "the" correlated measurement error structure, etc.

APPENDIX F--Derivation of Single Indicator Loadings and Measurement Error Variances

Werts, Linn and Jöreskog (1974) proposed that the latent variable reliability (ρ_X) of a of a unidimensional measure X (i.e., the measure has only one underlying latent variable) is given by

F1) $$\rho_X = \frac{L^2_X \mathrm{Var}(X)}{L^2_X \mathrm{Var}(X) + E_X} \ ,$$

where L_X is the sum of the loadings of the items in the measure X on their latent variable *X*, Var(*X*) is the (error disattenuated) variance of *X* (i.e., from a measurement model of *X*), and E_X is the sum of the measurement error variances of the items in the measure X as they load on their latent variable *X*. It is also well known that E_X is given by

F2) $$E_X = \mathrm{Var}(\mathrm{X})\ (1 - \rho_X)\ ,$$

where Var(X) is the (error attenuated) variance of X (e.g., obtained using SAS, SPSS, etc.). By solving Equation (F1) for L_X and substituting Equation (F2) into the result,

F3) $$L_X = [\mathrm{Var}(\mathrm{X})\ \rho_X / \mathrm{Var}(X)]^{1/2}$$

which becomes

F4) $$L_X = [\rho_X]^{1/2}$$

when Var(X) equals Var(*X*) (e.g., if *X*, and thus X, is standardized and its variance is equal to 1), or

F5) $$L_X \approx [\rho_X]^{1/2}$$

otherwise, where ≈ indicates "approximately equal to."

Finally, Anderson and Gerbing (1988) pointed out that for a unidimensional measure there is little practical difference between coefficient alpha (α) and latent variable reliability ρ.

Thus, for a single indicator specification of a standardized latent variable *X* (i.e., its variance is fixed at 1), its loading, L_X, is the square root of its latent variable reliability ρ (see Equation F1), and its measurement error variance, E_X, is 1 - ρ_X (see Equation F2).

These parameters can be estimated for a standardized latent variable *X* by substituting coefficient alpha reliability, α, into Equations (F2) and (F4), and for an unstandardized latent variable *X* its single indicator loading can be approximated using Equation (F5) and α , and its measurement error variance can be estimated using Equation F2.

Ping (2004) suggested summing maximum likelihood exploratory factor analysis (EFA) loadings for L_X, and using Equation F2 for the measurement error variance of a fully aggregated measure. Because EFA also produces an estimate of the Average Extracted Variance (AVE), the explained variance for a factor, the equation for the AVE of *X* (see Fornell & Larker, 1981)

$$\mathrm{AVE}_X = \frac{\Sigma(l_{xj\ (j=1,p)})^2\,\mathrm{Var}(X)}{\Sigma(l_{xj\ (j=1,p)})^2\,\mathrm{Var}(X) + E_X}$$

$$= \frac{\Sigma(l_{xj\ (j=1,p)})^2}{\Sigma(l_{xj\ (j=1,p)})^2 + E_X}\ ,$$

where l_{xj} is a loading, Σ is a sum (of squared loadings--Equation F1 involves the square of the sum), and Var(*X*) = 1, can be solved for the measurement error variance of a sum of indicators, E_X

$$\text{F6)}\qquad E_X = \frac{\Sigma(l_{xj\ (j=1,p)})^2\,(1 - \mathrm{AVE})}{\mathrm{AVE}}\ ,$$

where AVE is the explained variance of (unidimensional) factor containing the items in X.

Thus, an additional estimate of the loading of a summed indicator composed of unidimensional items is the sum of its EFA loadings, and an additional estimate of its measurement error variance is given by Equation (F6).

ENDNOTES

[1] In this case inconsistency is actually unacceptable consistency: the observed or input correlation between two indicators of the same latent variable is not acceptably numerically similar to the path analytic (Wright, 1934) product of the coefficients on the loading paths from their common latent variable.

[2] Using path analysis (Wright, 1934) the covariance of two unidimensional items x_1 and x_2 (i.e., items with only one underlying latent variable) implied by a model is the product of the path coefficients on their paths from their common latent variable (i.e., the product of their loadings), *plus* the product of the path coefficients due to correlated measurement error (i.e., $\lambda_1*\lambda_2 + 1*1*\mathrm{Cov}(\varepsilon_1,\varepsilon_2)$, where λ denotes loading, ε denotes measurement error, 1 is the implied path coefficient on the path between each measurement error and their respective x, and Cov denotes covariance). Because correlations between measurement errors are usually assumed to be zero, the covariance term is usually ignored.

[3] Content or face validity is usually established by qualitatively judging how well items match the conceptual definition of the target construct.

[4] A second-order construct has other constructs as its "indicators." For example in Dwyer and Oh's (1987) study of Environmental Munificence and Relationship Quality, the second-order construct Relationship Quality had the first-order constructs Satisfaction, Trust, and Minimal Opportunism as indicators (see Bagozzi, 1981b; Bagozzi and Heatherton, 1994; Gerbing and Anderson, 1984; Gerbing, Hamilton and Freeman, 1994; Hunter and Gerbing, 1982; Jöreskog, 1970; and Rindskopf and Rose, 1988 for accessible discussions of second-order constructs).

[5] Bagozzi and Heatherton (1994) also used a variation of this approach that did not use reliability loadings.

[6] For example, a model with just-identified latent variables (3 items per latent variable), and 5 latent variables requires 240 cases to produce at least two cases per input covariance matrix element. The same model with 5 summed indicators and 240 cases would have 16 cases available to compute each input covariance matrix element.

[7] The study details have been omitted to skirt matters such as conceptual definitions, hypotheses, etc. which were judged to be of minimal importance to the present purposes.

[8] Anderson and Gerbing (1984) suggested that GFI and AGFI may not be appropriate gauges of model-to-data fit in larger models. An RMSEA (Steiger, 1990) of .05 suggests close fit and values through .08 suggest acceptable fit--see Browne and Cudeck (1993); Jöreskog (1993).

[9] However, this indeterminacy could be remedied by reconvening an item-judging panel to judge the resulting measures and thus identify a weeded subset of items that best taps the conceptual definition of N.

[10] The variations of full aggregation (e.g., factor scores, LISREL 8 parameters, etc.) have not been formally investigated for structural coefficient bias and inefficiency, as far as we know, and the structural coefficient results from External Consistency Only and full aggregation should be compared for validation. A disagreement in nonsignificance (e.g., a structural coefficient varies between nonsignificance and significance with External Consistency Only and factor scores) should probably be judged nonsignificant.

[11] Other omissions were possible--e.g., omitting a Subset 2 item from the set of 18, etc., and there were 18 of these, some of which may have duplicated Subset 1.

[12] While these two itemizations had no items in common, equivalent behavior was observed with Subset 2 and Subset 1 or Subset 1a that did have common items. Parenthetically, the reliability of the antecedent latent variable was .86. However, similar behavior was observed for a latent variable with much lower reliability. Finally, there were other structural coefficients that were completely unaffected by changes in the itemizations of N.

[13] This process was actually repeated more than 180 times to check for errors because the results appeared to be cycling. Whether or not this process would have converged with this number of potential correlated measurement errors (171) is unknown.

[14] Correlating all measurement errors with ε_9 was not identified, so the correlation between ε_9 and ε_{13} was fixed at zero because its modification index (MI) was .0001, suggesting the correlation was nonsignificant. A MI in this case is approximately a Chi-Square statistic for freeing the correlation between ε_9 and ε_{13}--a MI of .0001 suggests the path coefficient on the correlation between ε_9 and ε_{13} would have a Chi-Square difference (from 0) of .0001 which is nonsignificant with a p-value of .992 and 1 degree of freedom.

[15] Using expectation algebra and the usual assumptions regarding latent variables and their errors of measurement, the variance of a sum of indicators x_i, $\mathrm{Var}(x_1+x_2+...+x_p) = \mathrm{Var}(\lambda_{x1}X + \varepsilon_{x1} + \lambda_{x2}X + \varepsilon_{x2} + ... + \lambda_{xp}X + \varepsilon_{xp}) = (\Sigma\lambda_{xj\ (j=1,p)})^2\mathrm{Var}(X) + \Sigma\mathrm{Var}(\varepsilon_{xj\ (j=1,p)})$, where λ is a loading, X is a latent variable, and ε is a measurement error.

[16] Because exploratory factor analysis assumes variances of 1, the loadings were re-scaled by dividing each loading by the maximum loading to allow for latent variable variances other than 1. This produces one loading equal to one and the other loadings in the customary .6 to .9 range.

[17] In retrospect, we should have at least numbered the scenarios by treatment.

ON ASSURING VALID MEASURES
FOR THEORETICAL MODELS USING SURVEY DATA

Robert A. Ping, Jr.
Associate Professor of Marketing
College of Business Administration
Wright State University
Dayton, OH 45435

(Revised: 2010, 2006--see end of paper)

ON ASSURING VALID MEASURES FOR THEORETICAL MODELS USING SURVEY DATA

This research critically reviews the process and procedures used in Marketing to assure valid and reliable measures for theoretical model tests involving unobserved variables and survey data, and it selectively suggests improvements. The review and suggestions are based on reviews of articles in the marketing literature, and the recent methods literature. This research also provides several perhaps needed explanations and examples, and is aimed at continuous improvement in theoretical model tests involving unobserved variables and survey data.

Based on the articles in our major journals, marketers generally agree that specifying and testing theoretical models using Unobserved Variables with multiple item measures of these unobserved variables and Survey Data (UV-SD model tests) involve six steps: *i*) defining constructs, *ii)* stating relationships among these constructs, *iii*) developing measures of the constructs, *iv*) gathering data, step *v*) validating the measures, and *vi*) validating the model (i.e., testing the stated relationships among the constructs). However, based on the articles reviewed (see Endnote 1 for these journals), there appears to be considerable latitude, and confusion in some cases, regarding how these six steps should be carried out for UV-SD model tests in Marketing.

For example in response to calls for increased psychometric attention to measures in theoretical model tests, reliability and validity now receive more attention in UV-SD model tests (e.g., Churchill, 1979; Churchill and Peter, 1984; Cote and Buckley, 1987, 1988; Heeler and Ray, 1972; Peter, 1979, 1981; Peter and Churchill, 1986). However, there were significant differences in what constitutes an adequate demonstration of measure reliability and validity in the articles reviewed. For example in some articles, steps *v)* (measure validation) and *vi)* (model validation) involved separate data sets. In other articles a single data set was used to validate both the measures and the model. Further, in some articles the reliabilities of measures used in previous studies were reassessed. In other articles reliabilities were assumed to be constants that, once assessed, should be invariant in subsequent studies. Similarly, in some articles many facets of validity for each measure were examined, even for previously used measures. In other articles few facets of measure validity were examined, and validities were assumed to be constants (i.e., once judged acceptably valid a measure was acceptably valid in subsequent studies).

Thus an objective of this research is to selectively identify areas for continuous improvement in step *v)*, measure validation. The research provides a selective review, albeit qualitative, of the UV-SD model testing practices of marketers in that step and the other steps as they pertain to step *v)*. It also provides selective discussions of errors of omission and commission in measure validation. For example, this research discusses the implications of reliability and facets of validity as sampling statistics with unknown sampling distributions. It suggests techniques such as easily executed experiments that could be used to pretest measures, and bootstrapping for reliabilities and facets of validity. The research also suggests an estimator of Average Variance Extracted (AVE) (Fornell and Larker, 1981) that does not rely on structural equation analysis (e.g., LISREL, EQS, AMOS, etc.). In addition, it suggests an alternative to omitting items in structural equation analysis to improve model-to-data fit, that should be especially useful for older measures established before structural equation analysis became popular.

MEASURE VALIDATION

Step *v)*, measure validation or demonstrating the adequacy of the study measures, appeared to be the least consistent of the six steps above (see Peter and Churchill 1986 for similar findings). Perhaps this was because there are several issues that should be addressed in validating measures. Measures should be shown to be unidimensional (having one underlying construct), consistent (fitting the model in structural equation analysis), reliable (comparatively free of measurement error), and valid (measuring what they should). Demonstrating validity has also been called measure validation (see Heeler and Ray, 1972). However, I will use the term measure validation to mean demonstrating measure unidimensionality, consistency (i.e., model-to-data fit), reliability, and validity.

While step *v)*, measure validation, is well-covered elsewhere, based on the articles reviewed it appears to merit a brief review. I begin with unidimensionality and consistency, then proceed to reliability and validity.

UNIDIMENSIONALITY

Assessing reliability usually assumes unidimensional measures (Bollen, 1989; Gerbing and Anderson, 1988; Hunter and Gerbing, 1982). However, coefficient alpha, the customary index of reliability in Marketing, underestimates the reliability of a multidimensional measure (Novick and Lewis, 1967). Thus, unidimensionality is actually required for the effective use of coefficient alpha (Heise and Bohrnstedt, 1970--see Hunter and Gerbing, 1982) (other indexes of reliability such as coefficient omega have been proposed for multidimensional measures -- see Heise and Bohrnstedt, 1970). Thus reliability of a measure, as it was typically assessed in the studies reviewed (i.e., using coefficient alpha), should be assessed after unidimensionality has been demonstrated (Gerbing and Anderson, 1988).

A unidimensional item or indicator has only one underlying construct, and a unidimensional measure consists of unidimensional items or indicators (Aker and Bagozzi, 1979; Anderson and Gerbing, 1988; Burt, 1973; Gerbing and Anderson, 1988; Hattie, 1985; Jöreskog, 1970 and 1971; McDonald, 1981). In the articles reviewed, unidimensionality was typically assumed in the specification of a model estimated with structural equation analysis. Perhaps this was because authors have stressed the need for unidimensionality in structural equation analysis models in order to separate measurement issues (i.e., the relationship between a construct and its observed variables or indicators) from model structural issues (i.e., the relationships or paths among constructs) (Anderson, Gerbing and Hunter, 1987; Anderson and Gerbing, 1988; Bentler, 1989; Bollen, 1989; Burt, 1976; Jöreskog, 1993) (however, see Kumar and Dillon, 1987a and 1987b for an alternative view). Separating measurement issues from model structural issues in structural equation analysis avoids interpretational confounding (Burt, 1976), the interaction of measurement and structure in structural equation models. In particular, an item or indicator x can be viewed as composed of variance due to its construct *X* and variance due to error, and thus

$$\mathrm{Var(x)} = \lambda^2\mathrm{Var}(X) + \mathrm{Var(e)}\ , \qquad (1$$

if *X* and e are independent, where Var denotes variance, λ or lambda is the path coefficient on the path connecting *X* with x (also called the loading of item x on *X*), and e is error. Intrepretational confounding in structural equation analysis means that changes in model structure (i.e., adding or deleting paths among constructs) can produce changes in the measurement parameter estimates of a construct (i.e., changes in item loadings, in measurement errors, and in construct variances). Thus, with interpretational confounding, changes in the structural equation model can affect the empirical meaning of a construct.

CONSISTENCY

Many criteria for demonstrating unidimensionality have been proposed (see Hattie, 1985). Perhaps in response to calls for more work in this area (e.g., Lord, 1980), Anderson and Gerbing (1982) proposed operationalizing unidimensionality using the structural equation analysis notions of internal and external consistency (also see Kenny, 1979; Lord and Novick, 1968; McDonald, 1981) (however see Kumar and Dillon, 1987a and 1987b for an alternative view).

Consistency has been defined as the structural equation model fitting the data (see Kenny, 1979). It is important because coefficient estimates from structural equation analysis may be meaningless unless the model adequately fits the data (Bollen, 1989; Jöreskog ,1993:297). As Anderson and Gerbing (1982) defined consistency, two indicators of *X*, x_1 and x_2, are internally consistent if the correlation between them is the same as the product of their correlations with their construct *X*. Similarly an indicator of *X* and an indicator of *Z*, x and z, are externally consistent if the correlation between x and z is the same as the product of three correlations: x with its construct *X*, z with its construct *Z*, and *X* with *Z*. Thus if *X* is internally and externally consistent, it is also unidimensional, and I will use the term consistent/unidimensional for Anderson and Gerbing's (1982) operationalization of consistency.

Anderson and Gerbing (1982) also proposed assessing consistency/unidimensionality with what they termed similarity coefficients (see Hunter, 1973; Tyron, 1935). The similarity coefficient for the items or indicators *a* and *b* in the same or different measures is the cosine of the angle between the vector of correlations of *a* with the other items in a study (including b), and the vector of correlations of *b* with the other study items (including *a*). Similar items have a small angle between their correlation vectors, and a cosine of this angle that is near one. Specifically, Anderson and Gerbing (1982) proposed that *a* and *b* have high internal consistency if their similarity coefficient is .8 or above. External consistency is suggested by items that cluster together in a matrix of sorted or ordered similarity coefficients (Anderson and Gerbing, 1982:458) (see Appendix B for an example).

Consistency/unidimensionality is also suggested by a structural equation model that fits the data when its constructs are specified as unidimensional (i.e., each observed variable or indicator is connected to only one construct). With consistency/unidimensionality there is little change in measurement parameter estimates (i.e., loadings and variances-- see Equation 1) between the measurement model and subsequent structural models (Anderson and Gerbing, 1988) (i.e., differences in second or third decimal digits only). Thus consistency/unidimensionality can also be suggested by showing little if any change in measurement parameters estimates between a full measurement model (i.e., one containing all the model constructs, and their indicators, with correlations among all the constructs) and the structural model (i.e., one that replaces certain correlations among the constructs with paths).

PROCEDURES FOR ATTAINING UNIDIMENSIONALITY AND CONSISTENCY

Procedures for attaining unidimensionality using exploratory (common) factor analysis are well known. However, procedures for obtaining consistent/unidimensional measures are less well documented. Procedures using ordered similarity coefficients are suggested in Anderson and Gerbing (1982:454), and Gerbing and Anderson (1988). The ordered similarity coefficients help identify inconsistent items. Alternatively, consistency/unidimensionality for constructs specified unidimensionally (i.e., each observed variable or indicator is "pointed to" by only one construct) can be attained using a procedure that has been in use for some time (see Dwyer and Oh, 1987; Kumar and Dillon, 1987b; Jöreskog, 1993) (however see Cattell, 1973 and 1978 for a dissenting view). The procedure involves estimating a single construct measurement model (i.e., one that specifies a single construct and its items) for each construct, then measurement models with pairs of constructs, etc., through estimating a full measurement model containing all the constructs. Items are omitted as required at each step to obtain adequate measurement model fit (and thus consistency/unidimensionality because the process begins with single construct measurement models) while maintaining content or face validity (content or face validity is discussed later and should be a serious concern in omitting items using any consistency improvement procedure). Standardized residuals, or specification searches (e.g., involving modification indices in LISREL or LMTEST in EQS) can also be used to suggest items to be omitted at each step to improve model-to-data fit.

However, these methods are not particularly efficient, and they may not always produce the largest consistent/unidimensional subset of indicators. Instead, partial derivatives of the likelihood function with respect to the error term of the indicators could be used to suggest inconsistent items (see Ping 1998a). This approach involves the examination of the matrix of these derivatives in a single construct measurement model. The item with the largest summed first derivatives without regard to sign that preserves the content or face validity of the measure is omitted. The matrix of first derivatives is then re estimated without the omitted item, and the process is repeated until the single construct measurement model fits the data (see Appendix A for an example of this procedure).

My experience with this procedure and real survey data sets is that it produces maximally internally consistent item subsets. The approach is similar to Saris, de Pijper and Zegwaart's (1987) and Sörbom's (1975) proposal to improve model-to-data fit using partial derivatives of the likelihood function with respect to fixed parameters (i.e., to suggest paths that could be freed, e.g., modification indices in LISREL). The internally consistent measures produced are frequently externally consistent. Nevertheless, the procedure could also be used on a full measurement model containing all the constructs specified unidimensionally (i.e., each observed variable or indicator is connected to only one construct). This full measurement model variant of the first derivative approach is useful if several study measures are inconsistent, because the most inconsistent item in each measure can be identified with a single measurement model.

COMMENTS ON UNIDIMENSIONALITY AND CONSISTENCY

Unidimensionality in the exploratory common factor analytic sense is required for coefficient alpha, and consistency/unidimensionality is required for structural equation analysis. Further, it is well known that the reliability of a measure is necessary for its validity. Thus, there is a sequence of steps in validating a measure: establish its consistency/unidimensionality for structural equation analysis, or establish its unidimensionality using maximum likelihood exploratory common factor analysis (i.e., not principal components factor analysis) for regression (however, see Endnote 2 for cautions about regression), then show its reliability, and finally its validity.

Unidimensionality in two and three item measures is difficult to demonstrate using exploratory or confirmatory factor analysis because these measures are under- or just determined. However, ordered similarity coefficients will gauge both internal and external consistency and thus unidimensionality using the criteria discussed above.

While Churchill and Peter (1984) found no effect on reliability when positively and negatively worded or reverse-polarity items are mixed in a measure, subsequent studies suggest that mixing positively and negatively worded items can adversely affect measure consistency/unidimensionality (see the citations in Herche and Engelland, 1996). If concern for acquiescence bias (see Ray, 1983) produces a measure with positively and negatively worded items that produces consistency/unidimensionality problems, inconsistent items might be retained as a second facet in a second-order construct (see Bagozzi 1981b for a similar situation) (second-order constructs are discussed later).

My experience with the above procedures in obtaining consistency/unidimensionality is that they are all tedious, especially the first derivative procedure. An alternative is to avoid consistency problems by summing one or more constructs' items and use regression (see Endnote 2 for cautions about regression), or use single indicator structural equation analysis (which will be discussed next). In addition, ordered similarity coefficients do not always suggest maximally consistent item clusters in survey data. Instead they usually suggest sufficiently consistent clusters of items that are also sufficiently reliable (see Appendix B for an example).

In survey data it is easy to show that unidimensionality obtained using maximum likelihood exploratory common factor analysis does not guarantee consistency/unidimensionality in the Anderson and Gerbing (1982) sense. Thus, consistency/unidimensionality is a stronger demonstration of unidimensionality than a single factor solution in maximum likelihood exploratory common factor analysis. Based on the articles reviewed and my own experience, there seems to be an upper bound for the number of items in a consistent/unidimensional measure of about six items (also see Bagozzi and Baumgartner, 1994 for a similar observation). Thus larger measures, especially older measures developed before structural equation analysis became popular, usually required extensive item omission to attain consistency/unidimensionality in the articles reviewed. While the resulting consistent/unidimensionality submeasures were invariably argued or implied to be content or face valid, they often seemed to be less so than the original full measures.

In fact, a common misconception in the reviewed articles that used structural equation analysis was that consistent measures are more desirable than less consistent fuller measures, especially older measures developed before structural equation analysis became popular. Many articles appeared to assume that older full measures were inherently flawed because they were typically inconsistent and required item omission to attain a consistent subset of items. Nevertheless, it could be argued that the full measures were frequently more desirable than the proposed more consistent reduced measures for reasons of face or content validity. Thus, I will discuss an alternative to item omission to attain consistency/unidimensionality in structural equation analysis.

Single Indicator Structural Equation Analysis Item omission to attain acceptable measurement model-to-data fit may not always be necessary in order to use structural equation analysis. In situations where it is desirable for reasons of face or content validity to use a unidimensional, in the exploratory common factor analysis sense, but less than consistent measure, the items in the measure could be summed and regression could be used to validate a UV-SD model (however see Endnote 2). Alternatively Kenny (1979) hinted at a procedure involving reliabilities that can be used with structural equation analysis to validate a UV-SD model. Variations of this procedure have been used elsewhere in the social sciences (see for example Williams and Hazer 1986 and the citations therein). This procedure involves summing the items in a measure that is unidimensional using maximum likelihood exploratory common factor analysis, then averaging them to provide a single indicator of the unobserved construct.

Because this single indicator specification is under determined, estimates of its loading and measurement error variance are required for structural equation analysis. The observed indicator x of an unobserved or latent variable X can be written $x = \lambda X + e$, where λ or lambda is the loading of x on X (i.e., the path coefficient on the path from the unobserved variable X to the observed variable x) and e is error. Thus, the loading Λ of the averaged indicator X (= $[x_1 + x_2 + ... + x_n]/n$) on X, is approximated by $\Sigma l_i/n$ in

$$\mathrm{X} = (x_1 + x_2 + ... + x_n)/n = (\lambda_{x1}X + e_{x1} + \lambda_{x2}X + e_{x2} + ... + \lambda_{xn}X + e_{xn})/n$$

$$\approx (\Sigma l_i X)/n + (\Sigma e_i)/n = \Lambda X + (\Sigma e_i)/n \;, \qquad (2$$

if X and e are independent, where l_i are the loadings of x_i on X from a maximum likelihood exploratory common factor analysis. (Note: in practice, each l_i should be divided by the largest l_i if $\mathrm{Var(X)} \neq 1$--see Appendix F.)

It is well known that an estimate of the measurement error variance of the averaged indicator X is Var(X)(1-ρ) (Kenny 1979), where Var(X) is the variance of X, and ρ is the latent variable reliability of X (discussed later-- see Equation 4). However, Anderson and Gerbing (1988) pointed out that for unidimensional measures there is little practical difference between coefficient alpha (α) and the latent variable reliability ρ. Thus for an unidimensional measure X, estimates of the loading of its averaged indicator and its measurement error are $\Lambda = \Sigma l_i/n$ and Var(X)(1-α) respectively, where Var(X) is the variance of the averaged indicator available in SAS, SPSS, etc. (see Appendix F for the details and an example).

This procedure can be simplified further. Authors have defined the reliability of a unidimensional indicator as the square of the loading between the indicator and its latent variable (see Bollen, 1989). Thus, the square root of α ($\sqrt{\alpha}$) could be substituted for Λ (Kenny, 1979) (see Appendix F). However, this substitution produces biased (i.e., understated) variance estimates for measures with highly variable loadings. While there is no hard and fast rule, $\sqrt{\alpha}$ should probably not be used with a measure that has items with reliabilities (i.e., $\sqrt{\alpha}$)

of less than .8 (i.e., λ's less than .64). (Note: √α is not biased if Var(X) is standardized to equal 1.)

Model-to-data Fit Consistency/unidimensionality can be established for models specified unidimensionally (i.e., each observed variable or indicator is "pointed to" by only one construct) using model-to-data fit (fit). Thus, one use of indices of fit is to suggest consistency/unidimensionality. Perhaps because there is no agreement on the appropriate index of fit (see Bollen and Long, 1993), multiple indices of fit were usually reported in the articles reviewed. The most commonly reported index of fit, chi-square, is a measure of exact fit (Browne and Cudeck, 1993). However it rejects model-to-data fit as the number of cases increases (Hoelter, 1983), and additional fit statistics such as Goodness of Fit Index (GFI) and Adjusted Goodness of Fit Index (AGFI) were typically reported in the articles reviewed. However, GFI and AGFI decline as model complexity increases (i.e., more observed variables or more constructs), and they may be inappropriate for more complex models (Anderson and Gerbing, 1984).

In addition to chi-square, GFI, and AGFI, the articles reviewed variously reported standardized residuals, comparative fit index (CFI), and root mean square error of approximation (RMSEA), among other indices of fit. Because there is also no agreement on an appropriate set of fit indices, I will simply note that standardized residuals were reported increasingly less frequently over time in the articles reviewed. Bentler's (1990) CFI appeared to be growing in popularity and was frequently reported in the recent articles, as was Steiger's (1990) RMSEA (possibly because it appears to have Jöreskog's 1993 endorsement) (see Endnote 3 for more on CFI and RMSEA).

RELIABILITY

Unfortunately the term consistency has been used in connection with reliability (see for example DeVellis, 1991:25). In fact there has been considerable confusion over reliability and consistency (see Hattie, 1985). After discussing reliability I will discuss the distinctness of reliability from consistency as Anderson and Gerbing (1982) have defined it.

Measure reliability was usually reported in the articles reviewed. The reliability of a measure is suggested by agreement of two efforts to measure its construct using maximally similar methods (Campbell and Fiske, 1959). Thus it is frequently characterized as the "repeatability" of a measure, and types of reliability include a measure's stability over time or subjects (see Bollen, 1989; Nunnally, 1978). It is also described in terms of the amount of random error in a measure (Lord and Novick, 1968; see Bollen, 1989 and Nunnally, 1978). For example, the variance of an indicator x of a construct *X* could be viewed as composed of variance due to its construct *X* and variance due to error (see Equation 1). As a result, The reliability ρ of a consistent/unidimensional item x has been operationalized as the ratio of its variance due to its construct, $\lambda^2\mathrm{Var}(X)$, and the total variance of x,

$$\rho_x = \frac{\lambda^2\mathrm{Var}(X)}{\mathrm{Var}(x)} = \frac{\lambda^2\mathrm{Var}(X)}{\lambda^2\mathrm{Var}(X) + \mathrm{Var}(e)} , \qquad (3$$

if *X* and e are independent, where Var(*X*) is the disattenuated (measurement-error-free) variance of *X* available in a structural equation measurement model (Werts, Linn and Jöreskog, 1974).

While there have been many proposals for assessing reliability (see Hattie, 1985; Nunnally, 1978), coefficient alpha (Cronbach, 1951) is generally preferred (Peter, 1979) because it does not depend on the assumptions required of other indices of reliability (see Bollen, 1989). However, coefficient alpha assumes that its items are perfectly correlated with their underlying construct (i.e., measured without error) (see Bollen, 1989). Because this assumption is almost always unreasonable in practice, coefficient alpha underestimates reliability (see Smith, 1974).

There have been several proposals for computing reliability of items that are measured with error (see Gerbing and Anderson 1988 for a summary). The most frequently used formula is due to Werts, Linn and Jöreskog (1974) (see Bagozzi, 1980b; Bollen, 1989; Dillon and Goldstein, 1984; Fornell and Larker, 1981). This Latent Variable Reliability of a measure X, with indicators (items) x_1, x_2, ... , x_n, is given by,

$$\rho_X = \frac{(\Sigma\lambda_i)^2\mathrm{Var}(X)}{(\Sigma\lambda_i)^2\mathrm{Var}(X)+\Sigma\mathrm{Var}(e_i)} , \qquad (4$$

where λ_i is the loading of x_i on *X*, e_i is the error term for x_i, Var(*X*) is the disattenuated (measurement error free) variance of *X* (i.e., available in a structural equation measurement model), and Σ denotes a sum.

However as previously mentioned, Gerbing and Anderson (1988) pointed out that for unidimensional measures there is little practical difference between coefficient alpha and Latent Variable Reliability. Thus to demonstrate reliability, it may be sufficient to report coefficient alpha because at worst it provides a conservative estimate of reliability.

Based on the articles reviewed, it is important to note that reliability and consistency as it was just discussed are distinct notions (Green, Lissitz and Mulaik, 1977). An item could be consistent with other items but unreliable because of measurement error (see Equation 3). It is also easy to show using survey data that maximizing reliability (see Churchill, 1979) may not maximize consistency, and that reliable measures may not fit the data well using structural equation analysis (see Gerbing and Anderson, 1988 and Appendix A).

VALIDITY

There was considerable variation in the demonstrations of validity among the articles reviewed. This may be because methods authors do not all agree on what constitutes an adequate demonstration of validity. Item validity is how well an item measures what it should, and a valid measure consists of valid items. Validity is important because theoretical constructs are not observable, and relationships among unobservable constructs are tested indirectly via observed variables (Jöreskog, 1993; see Bagozzi, 1984). Thus validity reflects how well a measure reflects its unobservable construct. It is established using relationships between observed variables and their unobserved variable, and observed variables' relationships with other sets of observed variables (Jöreskog, 1993).

The following discussion assumes unidimensional and reliable measures. While methods author do not agree on a maximal set of validity tests, validity should be gauged using at least the following criteria: content or face validity (how well items match their conceptual definition), criterion validity (measure correspondence with other known valid and reliable measures of the same construct), and construct validity (measure correspondences with other constructs are consistent with theoretically derived predictions) (e.g., Bollen, 1989; DeVellis, 1991; Nunnally, 1978). Overall measure validity is then qualitatively assessed considering its reliability and then its performance over this minimal set of validity criteria.

The above terms for validity criteria are from the psychological and sociological literatures. However, other labels have been used, especially in Marketing. For example, content validity has been called face or consensus validity (see Heeler and Ray, 1972). Construct validity was used by Peter (1981) for content validity. Construct validity has been called nomological validity (see Peter, 1981). Trait validity has been used for a combination of reliability, and convergent and discriminant validity (Campbell, 1960). Finally, demonstrating convergent and discriminant validity has been called measure validation (see Heeler and Ray, 1972).

Content or Face Validity Content or face validity was not consistently demonstrated in the articles reviewed. Conceptual definitions, the definitions of the constructs comprising the UV-SD model, are required to provide conceptual meaning for the constructs in the model, and they are the basis for gauging the construct or face validity for these constructs. In the articles reviewed, conceptual definitions were not consistently given, especially for previously measured concepts. In fact, many articles appeared to assume that because a measure had been judged content or face valid in a previous article, all subsequent readers would accept the measure as content or face valid. However, in some cases it could be argued that the content or face validity of a measure was still an open matter, even though it had been judged to be content or face valid in a previous article.

In addition, conceptual definitions were not always stated for new measures. Further, item judging was frequently not discussed, and in many cases the full measure's items were not reported. It is difficult to imagine that these matters were not important during the review of these articles. Thus, while content validity may have been an important matter for reviewers, many articles left the impression that authors or editors consider these matters unimportant for journal readers.

Thus conceptual definitions should be clearly stated for each construct to enable readers to judge the content or face validity of measures of the constructs, even for previously used measures. In addition, care should be taken not to sacrifice evidence of content validity in the name of article space, for example.

Criterion Validity Criterion validity concerns the correspondence of a measure with a criterion measure, a known and, preferably, standard measure of the same concept. It is typically established using correlations.

However, there are no guidelines for adequate correlation between a measure and a criterion variable. In addition, for a new construct or a measure of an existing construct in a new context, a criterion measure may not be available. Perhaps for this latter reason criterion validity was rarely assessed in the articles reviewed.

Nevertheless, it was easy to wonder why a new measure of a previously measured construct was necessary in many cases, and how well a proposed measure of a previously measured construct would have fared in an assessment of criterion validity. Thus, for new measures of previously measured constructs criterion validity should be assessed and reported to improve the demonstration of a new measure's validity.

Construct Validity Construct validity is concerned in part with a measure's correspondence with other (i.e., different, non criterion) constructs. To begin to suggest construct validity, measures of other constructs should be valid and reliable, and their correspondences with the target measure should be theoretically sound. When it was considered in the articles reviewed, construct validity was typically suggested using correlations. The correlations with a target measure and their plausibility (i.e., their significance, direction and magnitude) were argued to support or undermine its construct validity.

Convergent and Discriminant Validity Convergent and discriminant validity are Campbell and Fiske's (1959) notions involving the measurement of multiple traits or constructs with multiple methods, and they are usually considered to be facets of construct validity in the social sciences. Convergent measures are highly correspondent (e.g., correlated) across different methods such as a survey and an experiment (such as scenario analysis-- discussed later). Discriminant measures are less correspondent with measures of other concepts than they are internally convergent.

Procedures for demonstrating convergent and discriminant validity using multiple traits and multiple methods is well documented (e.g., Bollen, 1989; Heeler and Ray, 1972). However, convergent and discriminant validity were seldom assessed in the articles reviewed as Campbell and Fiske (1959) intended. Perhaps because traits or constructs were typically measured with one method, reliability was frequently substituted for convergent validity, and measure distinctness (i.e., low correlations with other measures) was substituted for discriminant validity. However, while Nunnally (1978) suggested that a .7 or higher reliability implies convergent validity, measures with reliabilities above .85 can contain more than 50% error variance (see Appendix A). Thus measures with .7 or higher reliability may not be judged convergent valid because they contain less variance due to their construct than variance due to error.

Average Variance Extracted Perhaps for this reason, a statistic involving the percentage error variance in a measure, Average Variance Extracted (AVE) (Fornell and Larker 1981), was occasionally used to gauge convergent validity in the typically mono method studies reviewed. To explain AVE, the variance of a measure can be expressed as,

$$\mathrm{Var}(x_1+...+x_n) \;=\; \mathrm{Var}(\lambda_1 X+e_1+...+\lambda_n X+e_n) \;=\; (\Sigma\lambda_i^2)\mathrm{Var}(X)+\Sigma\mathrm{Var}(e_i), \qquad (5$$

if X and e are independent, where λ_i is the loading of the indicator x_i on the latent variable X., Var(X) is the disattenuated (error free) variance of X, and e_i is the measurement error of x_i. AVE is given by,

$$\mathrm{AVE}_X \;=\; \frac{(\Sigma\lambda_i^2)\mathrm{Var}(X)}{(\Sigma\lambda_i^2)\mathrm{Var}(X)+\Sigma\mathrm{Var}(e_i)}, \qquad (6$$

where Σ indicates a sum (see Endnote 4 more). The result is the percentage of the total variance of a measure (see Equation 5) represented or extracted by the variance due to the construct, $\lambda_1^2\mathrm{Var}(X) + ... + \lambda_n^2\mathrm{Var}(X) = (\Sigma\lambda_i^2)\mathrm{Var}(X)$. AVE ranges from 0 to 1, and Fornell and Larker (1981) suggested adequately convergent valid measures should contain less than 50% error variance (i.e., AVE should be .5 or above) (also see Dillon and Goldstein, 1984, and see Appendix E for an example).

Because acceptably reliable measures can contain more than 50% error (e.g., X in Appendix A), in UV-SD model tests a measure's reliability should probably be higher than Nunnally's (1978) suggestion of .7 to avoid a low AVE. While there is no firm rule, measure reliability should probably be .8 or more, to avoid these difficulties. However, a more precise alternative to reliability as a gauge of convergent validity would be an AVE of .5 or above. Thus, adequate convergent validity could be suggested by reliabilities of .8 or higher, and demonstrated by an AVE above .5.

Discriminant validity was typically established in the articles reviewed by using correlations when it was demonstrated. Although there is no firm rule, correlations with other measures below |.7| were usually accepted as evidence of measure distinctness and thus discriminant validity. Larger correlations were occasionally tested by examining the confidence intervals of correlations to see if they included 1 (see Anderson and Gerbing, 1988). They were also infrequently tested by using a single degree of freedom test that compares two structural equation measurement models, one with the target correlation fixed at 1, and a second with this correlation free (see Bagozzi and Phillips, 1982). If the difference in resulting chi-squares is significant, this suggests the correlation is not 1, and this implies the constructs are distinct, and it provides evidence of discriminant validity. (Note: this test and the correlation confidence interval test are untrustworthy for gauging discriminant validity--see "Is there any way to improve Average Variance Extracted (AVE) in a Latent Variable (LV) X?" on this website.)

Occasionally AVE was used to gauge discriminant validity. If the squared correlation between constructs (r^2) is less than either of their individual AVE's, this suggests the constructs each have more error free (extracted) variance than variance shared with other constructs (r^2). In different words, they are more internally correlated than they are with other constructs. This in turn suggests discriminant validity.

COMMENTS ON MEASURE VALIDATION

New measures frequently seemed to be underdeveloped in the articles reviewed. For example, new measure development details were not always reported. Thus it appeared that recommended procedures such as item judging, focus groups, etc. (Churchill, 1979, see Calder, 1977) were not always used to develop new measures. Several data sets should also be used to gauge the reliability and facets of the validity of measures (Campbell and Fiske, 1959; see Churchill, 1979). However, measure validation studies were seldom discussed. In some cases measure validation was abbreviated in a typically small pretest that was briefly summarized in the article, and the reliability and validity of the study measures was gauged using the final test data (i.e., the data used to test the proposed model).

Scenario Analysis While reliability and validity should always be confirmed using the final test data, care should be taken to conduct and adequately report measure validation studies, or the study results should be termed preliminary because the measures have received minimal testing. Although not reported in the articles reviewed, scenario analysis has been used elsewhere in the social sciences, and it could be used to produce data sets for preliminary measure validation (see Ping, 1998b). Scenario analysis is an experiment in which subjects (typically students) read written scenarios in which they are asked to imagine they are the subjects of an experiment in which variables are verbally manipulated. Then these subjects are asked to complete a questionnaire containing the study measures (see Appendix C). The results of scenario analysis have been reported to be similar enough to those from surveys to suggest that scenario analysis may be useful in new measure development and the verification of existing measures (see for example Rusbult, Farrell, Rogers and Mainous, 1988, and Appendix D).

Previously-used Measures As discussed earlier, many of the descriptions of previously-used measures in the articles reviewed were incomplete. For example, while the source of a previously-used measure was invariably given, the reliability and validity of these measures in previous studies were frequently left for the reader to find elsewhere. More important, articles frequently assumed that a demonstration of adequate reliability and validity of a measure in a previous study implied its reliability and validity in subsequent studies. Reliability and facets of validity such as construct, convergent, and discriminant validity are demonstrated using sample statistics (e.g., coefficient alpha for reliability, correlations for construct and discriminant validities, and AVE for convergent and discriminant validity) that vary from sample to sample (see Peter and Churchill, 1986). In addition, reliability and AVE have unknown sampling distributions, so they cannot be generalized beyond the study sample without additional samples (see Endnote 5 for generalizibility theory). Further, the content or face validity of previously-used measures occasionally seemed questionable, and some had actually performed marginally (i.e., exhibited low reliability or validity) in previous studies. Thus, care should be taken to show that previously-used measures are valid and reliable in the study being reported, and that they have been consistently so.

Reliability And Average Variance Extracted Reliability and Average Variance Extracted (AVE) are linked, but not always closely. While reliability is always larger than AVE (see Equations 4 and 5), a highly reliable measure can have an unacceptable AVE (e.g., in Appendix A, X has reliabilities of .81 to .86 and

AVE's of .5 or below). As Appendix A also suggests, it is possible to decrease reliability but increase AVE (see Tables A2 and A3). Thus, omitting items to improve reliability or consistency can either improve or degrade AVE. Because omitting unreliable or inconsistent items may also undermine content validity, the final itemization of a measure can be a trade off among consistency/unidimensionality, reliability, AVE, and content or face validity.

Interactions and Quadratics In experiments with categorical independent variables (e.g., experiments analyzed with ANOVA), interactions (e.g., XZ in $Y = b_0 + b_1X + b_2Z + b_3XZ + b_4XX$) and quadratics (e.g., XX) are routinely investigated to help interpret significant main effects (i.e., the X-Y and Z-Y effects). However, interactions and quadratics were seldom investigated in the UV-SD model tests reviewed, even when theory seemed to suggest their existence. Although not reported in the few articles that did investigate interactions or quadratics, the reliability of these variables can be low. The reliability of XZ, for example, is

$$\rho_{XZ} = \frac{r_{XZ}^2 + \rho_X\rho_Z}{r_{XZ}^2 + 1} \quad , \qquad (7$$

where ρ denotes reliability and r_{XZ}^2 is the correlation of X and Z. Similarly, the reliability of XX is

$$\begin{aligned} \rho_{XX} &= \frac{\mathrm{Var}(X_TX_T)}{\mathrm{Var}(XX)} \\ &= \frac{2\mathrm{Var}^2(X_T)}{2\mathrm{Var}^2(X)} \\ &= (\rho_X)^2 \qquad (8 \end{aligned}$$

(Busemeyer and Jones, 1983) (Note: this result was incorrectly stated in the final *JBR* article). Thus, the reliability of an interaction or quadratic is approximately the product of the reliabilities of their constituent variables X and Z (see Endnote 6 for more). As a result, the reliabilities of the constituent variables that comprise an interaction or quadratic should in general be high. Further, Equations 7 and 8 do not produce the same values as the formula for coefficient alpha. Thus the SAS, SPSS, etc. programs that determining reliability can not be used for determining the reliability of an interaction or a quadratic.

The validity of interactions and quadratics was not considered in the articles examined. Specifically, content (there is an interaction estimation procedure that suggests dropping items-- see Jaccard and Wan, 1995), convergent, and discriminant validity should be considered for these variables. Interactions and quadratics are unavoidably correlated with their constituent variables, and thus they should be shown to be distinct from them (i.e., they should be shown to be discriminant valid). In addition, since the convergent validity of an interaction or quadratic measured using AVE is always less than its reliability, the convergent validity of an interaction or quadratic could be quite low and it should be reported.

Second Order Constructs There are several types of constructs in the social sciences, including the familiar first-order construct, and based on the articles examined, the somewhat less familiar second-order construct. A *first-order* construct has observed variables (i.e., measure items) as indicators of the construct. The relationship between indicators and a first order construct typically assumes the construct "drives" the indicators (i.e., the indicators are observable instances or manifestations of their unobservable construct, and a diagram of the construct and its indicators would show the construct specified or connected to the indicators with arrows from the construct to the indicators-- a reflexive relationship, see Bagozzi, 1980b and 1984). However, indicators can also "drive" their construct (i.e., the indicators define the construct, and a diagram of the construct and its indicators would show the indicators connected to the construct with arrows *from* the indicators to the construct-- a formative relationship, see Fornell and Bookstein, 1982). Formative constructs were seldom seen in the articles reviewed. Because formative constructs are not unobserved variables I will not discuss them further.

Occasionally a *second-order* construct was reported in the articles reviewed. These are constructs with other constructs as their indicators. For example in Dwyer and Oh's (1987) study of environmental munificence and relationship quality in interfirm relationships, the second-order construct relationship quality had the first-order constructs satisfaction, trust, and minimal opportunism as indicators (see Bagozzi, 1981; Bagozzi and Heatherton, 1994; Gerbing and Anderson, 1984; Gerbing, Hamilton and Freeman, 1994; Hunter and Gerbing, 1982; Jöreskog, 1970; and Rindskopf and Rose, 1988 for discussions of second-order constructs). Each first order construct in turn had their respective observed indicators. Presumably specifying the second order construct relationship quality with first order constructs simplified the model, yet it provided a richer model of the consequences of environmental munificence.

As the Dwyer and Oh example suggests, a second-order construct can be used to combine several related constructs into a higher-order construct using structural equation analysis. Thus a second-order construct could be used as an alternative to omitting items in a multidimensional construct (see Gerbing, Hamilton and Freeman, 1994 for examples). A second-order construct could also be used as an alternative to omitting inconsistent items, especially in older measures. If the omitted items are consistent among themselves, they could be specified as the second, third, etc. facet in a second order construct.

However, the reliability and validity of these second-order constructs were not reported. The coefficient alpha of these variables is computed using a dissattenuated (error-free) covariance matrix of the first-order constructs, or using the error variances (ζ's) and loadings (β's) of the first-order constructs on the second-order construct in a second-order measurement model in place of λ's and Var(e)'s Equation 4.

Similarly, the content or face validity and construct validities of second-order constructs should be reported. In this case, validity is demonstrated first by demonstrating valid first-order constructs, then by demonstrating the validity of the second-order construct with the first order constructs as indicators. The content validity of a second order construct is demonstrated as it is for first order constructs. However, the first order constructs should be viewed as the indicators of the second-order construct. Construct validity is suggested by plausible correlations of the second-order construct with the other study variables, while convergent validity could be suggested by an AVE for the second-order construct that is greater than .5. This AVE should be calculated using the loadings (β's) and measurement error variances (ζ's) of the first-order constructs on the second-order construct in a measurement model, or using Endnote 4 (see Endnote 7 for more).

Bootstrapping Although it was not reported in the articles reviewed, bootstrapping (Efron, 1981) could be used to produce confidence intervals for reliability and facets of validity, and thus provide them with a type of generalizability using the UV-SD model test data. Bootstrapping has been suggested to estimate standard errors (see Efron, 1981), and it has also been suggested to improve the asymptotic correctness of a sample covariance matrix (Jöreskog and Sörbom, 1996a:173, 185; see Bentler, 1989:76). It is accomplished by averaging the statistics of interest (in this case reliability, AVE, the disattenuated correlation matrix of the constructs, etc.) that result from taking subsamples of the available cases (e.g., a hundred subsamples each with 10% of the cases randomly deleted). The resulting bootstrapped (i.e., averaged) reliability, for example, and the square root of the variance of the reliability estimates could then be used to gauge a measure's reliability across multiple studies of the same population. A 95% confidence interval for the bootstrapped (average) reliability of a measure, ρ_{avg}, for example, would be $\rho_{avg} \pm 2(\mathrm{Var}_\rho)^{1/2}$ where $(\mathrm{Var}_\rho)^{1/2}$ is the square root of the variance of the reliabilities generated by the bootstrap procedure (the square root of which is an estimate of the standard error of ρ_{avg}). Multi-sample reliability could then be gauged by inspecting this confidence interval to see if it extended below .7. If it did, this would undermine the reliability of the target measure. Similarly, a 95% confidence interval for the bootstrapped (i.e., averaged) AVE, AVE_{avg}, would involve AVE_{avg} and $\mathrm{Var}_{\mathrm{AVE}}$., and an AVE_{avg} - $2(\mathrm{Var}_{ave})^{1/2}$ value less than .5 would undermine the convergent validity of the target measure. (Note: bootstrap results should be used with caution--see "Why are reviewers complaining about the use of PLS in my paper?" on this website.)

A bootstrapped (averaged) correlation matrix of the study constructs could also be used as an improved correlation matrix (i.e., more likely to be asymptotically, or large sample, correct) for gauging construct validity. In addition, if the confidence intervals for $(\mathrm{AVE}_X)^{1/2}$ or $(\mathrm{AVE}_Z)^{1/2}$ and the bootstrapped correlation between X and Z overlapped, that would undermine discriminant validity. (Again note that ibootstrap results should be used with caution--see "Why are reviewers complaining about the use of PLS in my paper?" on this website.)

IN CONCLUSION

Based on the articles reviewed it was difficult to escape the conclusion that reliability and validity in UV-SD model tests could be improved by simply following well-known procedures for this purpose (e.g., Churchill 1979). For example, an examination of the equations for reliability and convergent and discriminant validity suggest that difficulties with reliability or these facets of validity could be viewed as a result of insufficient error-free variance. Thus procedures for improving the reliability and validity of a measure should include increasing the error-free variance of its items, and increasing its items' loadings or correlations with its unobserved construct. In particular, increasing the number of item scale points, wording item stems in the language of the study population, and pretesting the study protocol (e.g., cover letter, questionnaire, etc.) deserve emphasis because they are easily implemented and particularly effective. Specifically, to increase construct variance and reduce measurement error variance, the number of scale points could be increased by replacing the ubiquitous five point Likert scale with a seven-point Likert scale, a ten-point rating scale, etc. (see Churchill and Peter, 1984).

Churchill's (1979) suggestion of using focus groups in item development is extensively used in applied marketing research to improve itemization and thus measurement error. Researchers in this venue believe that one or more small and convenient focus groups from the study population will yield important "instances" of observable sentiments and behaviors pertaining to study constructs that can reduce the guesswork in identifying valid items for a new or revised measure. In addition, these focus groups can reveal the specific language the study population uses to communicate regarding these constructs. This information is then used to improve the phrasing of item stems, and thus reduce measurement error.

Similarly, even rudimentary pretests should be effective in reducing measurement error. For example, administering the survey protocol (e.g., cover letter, questionnaire, etc.) to as few as one subject from the study population, then discussing their responses with them can be effective in reducing measurement error (see Dillon, Maden and Firtle, 1987:375).

Based on the articles reviewed, it may not be widely understood that reliability and facets of validity are actually sample-based statistics. Thus the reliability and facets of the validity of each study measure, including previously used measures, will vary across samples. Specifically, reliability and facets of validity cannot be generalized without additional samples because their sampling distributions are unknown. Thus the reliability and validity of each study measure should be evaluated and reported in a UV-SD model test, regardless of whether or not it has been used previously.

Because content or face validity in UV-SD model tests is subjectively gauged, not only by writers and reviewers but also by readers after the study is published, conceptual definitions, items, and measure development details should be reported in published articles so that subsequent readers can judge the content or face validity of each measure. Similarly, reliabilities and average extracted variances (AVE's), and a full correlation matrix for the constructs should also be reported so that construct, convergent and discriminant validity can be confirmed by readers.

A single data set was frequently used to validate both the measures and the model in the UV-SD model tests reviewed. Scenario analysis was suggested to provide additional data sets that could be used to preliminarily evaluate new and previously used measures. Bootstrapping the UV-SD model test data was suggested to gauge the generalizability of the reliability and facets of validity of the study measures. However, my preliminary experience with bootstrapped confidence intervals and sample sizes of 200 or more suggest that a measure with an observed reliability above .75 is unlikely to have a confidence interval that extends below .7, and that a measure with an observed average variance extracted (AVE) above .55 is unlikely to have a confidence interval that extends below .5.

However, it is possible for a measure to have a reliability above .8 yet have an AVE below .5 (see Appendix A). Thus new measures should have reliabilities and AVE's above .80 and .55, respectively, to improve the likelihood that their population values for AVE are above .5.

For structural equation analysis consistency is required to attain model-to-data fit and avoid interpretational confounding. Consistency can be attained using full measurement models and specification searches (e.g., modification indices in LISREL and LMTEST in EQS) to identify items that load significantly on multiple constructs. It can also be attained by omitting measure items that either do not cluster together in an ordered similarity coefficient matrix of all the measures, or have a large summed first derivative using a full measurement model.

However, omitting items to attain consistency can affect content validity, and item deletion should be done with care. The current practice of omitting items in older well-established measures to attain consistency/unidimensionality in structural equation analysis may be ill advised because it can reduce content or face validity. Alternatives to omitting items in a measure to attain consistency include summing them and using a single averaged indicator and regression (however see Endnote 2). They also include using structural equation analysis with a loading and a measurement error that are functions of the communalities or reliability of the items.

When structural equation analysis is used, model-to-data fit and parameter estimates (i.e., loadings, measurement errors, and construct variances and covariances) from a full measurement model (i.e., containing all the study constructs) should be reported so readers can verify the consistency/unidimensionality, reliability and AVE of the study constructs.

Needed Research It would be helpful to have additional insights into several aspects of reliability and validity. For example, the sampling distributions of the popular reliability coefficients and the average variance extracted (AVE) statistic are unknown. Approaches to providing approximate confidence intervals for these statistics include curve fitting an approximate distribution using Monte Carlo simulations. The results of Monte Carlo simulations involving data sets with different levels of measurement error could also be used to determine ranges of confidence intervals at differing reliabilities and/or AVE.

Similarly, reliability and AVE are related, and it is possible to have a measure with acceptable reliability yet unacceptable AVE using existing cutoffs for the acceptability of these statistics. This raises questions such as, should the limits of acceptable reliability and AVE be reconciled? If so, how? I suggested a conservative approach of raising the reliability cutoff to correspond to the AVE cutoff of .5, but what are the effects of lowering the AVE cutoff to correspond to the reliability cutoff of .7?

In addition, the assessment of reliability and validity in interactions and quadratics needs more work. For example, Busemeyer and Jones (1983) derived the formula for the reliability of an interaction (see Equation 7), yet Equation 7 produces a different result from the latent variable (LV) reliability equation

(Equation 4). Both approaches to determining reliability are plausible, yet the Busemeyer and Jones (1983) Equation 7 results are typically much larger than the LV reliability Equation 4 results. Since multiple approaches to determining reliability are practically equivalent for unidimensional first order latent variables, why do they typically produce different results in interactions and quadratics?

Similarly, if the Busemeyer and Jones (1983) Equation 7 formula is correct for the reliability of interactions and quadratics, and the LV reliability Equation 4 formula is not, is the Equation 6 formula for the AVE of an interaction correct? In this case would a modification of the Busemeyer and Jones (1983) Equation 7 formula be a more appropriate assessment of AVE for interactions and quadratics?

ENDNOTES

1. I reviewed UV-SD model tests in major marketing journals. The journals included *Journal of Marketing*, the *Journal of Marketing Research*, *Marketing Science*, the *Journal of Consumer Research*, the *Journal of the Academy of Marketing Science*, the *Journal of Retailing*, the *Journal of Personal Selling and Sales Management*, and the *Journal of Business Research* from 1980 to the present. I also reviewed the recent methods literature in the *Journal of Marketing Research*, the *Psychological Bulletin/Psychological Methods*, *Psychometrika*, *Multivariate Behavioral Research*, *Sociological Methodology*, and *Sociological Methods and Research*.

2. However, authors have warned against the used of regression with variables measured with error (see Bohrnstedt and Carter, 1971; Rock, Werts, Linn and Jöreskog, 1977; Warren, White and Fuller, 1974; and demonstrations in Cohen and Cohen, 1983), and from 1980 to the present regression was increasingly less frequently used in the articles reviewed.

3. CFI compares model fit to the fit of a null or independence baseline model (i.e., one in which the observed variables are composed entirely of measurement error). It typically varies between 0 and 1, and values .90 or above are considered indicative of adequate fit (see McClelland and Judd, 1993).

RMSEA has been suggested as a third indicator of fit (see Jöreskog, 1993), possibly because of the potential inappropriateness of chi-square, GFI and AGFI, and criticisms of CFI's all-error baseline model (see Bollen and Long, 1993). An RMSEA below .05 suggests close fit, while values up to .08 suggest acceptable fit (Browne and Cudeck, 1993; see Jöreskog 1993).

4. AVE must be manually calculated, and its parameters are available in structural equation modeling. AVE can be approximated using estimates of the Equation 6 parameters available from SPSS, SAS, etc. In Equation 6, $\Sigma\lambda_i^2$ is approximated by the sum of squares of the loadings in a maximum likelihood exploratory common factor analysis (i.e., the sum of the communalities or the eigenvalue of the items). Var(X) can be set to one, and ΣVar(e_i) is approximated by n - $\Sigma\lambda_i^2$. Thus, AVE is approximately the explained variance of the items in a maximum likelihood exploratory common factor analysis. Our experience is that for an unidimensional measure there is little practical difference between an AVE calculated using Equation 6 and an AVE equaling the percent explained variance in a maximum likelihood exploratory common factor analysis.

5. Generalizability theory (Cronbach, Gleser, Nada and Rajaratnam, 1972) can be used to address this problem for reliability, although it has not been widely used in the social sciences and was not used in any of the articles reviewed. Similarly bootstrapping, which is discussed later, could be used for reliability, AVE, etc. Both of these approaches, however require additional samples.

6. The exact reliability of an interaction is .002 to .228 larger than the product of the reliabilities of the constituent variables across constituent reliabilities of .7 to .9 and constituent correlations of .1 to .9-- the largest difference is for a correlation of .9 and constituent reliabilities of .7 each. The exact reliability of a quadratic is .095 to .225 larger than the square of the reliability of the constituent variable across constituent reliabilities of .7 to .9-- the largest difference is for a correlation of .9 and a constituent reliability of .7.

7. The indicator first-order constructs may not be discriminant valid, but this is not unusual in second order constructs.

REFERENCES

Aiken, Leona S. and Stephen G. West (1991), *Multiple Regression: Testing and Interpreting Interactions*. Newbury Park, CA: SAGE Publications.

Aker, D. A., and R. P. Bagozzi (1979), Unobservable Variables in Structural Equation Models with an Application in Industrial Selling. *Journal of Marketing Research*, 16, 147-158.

Anderson, James C. and David W. Gerbing (1982), Some Methods for Respecifying Measurement Models to Obtain Unidimensional Construct Measurement. *Journal of Marketing Research*, 19 (November), 453-60.

_______ and David W. Gerbing (1984), The Effect of Sampling Error on Convergence, Improper Solutions, and Goodness of Fit Indices for Maximum Likelihood Confirmatory Factor Analysis. *Psychometrika*, 49, 155-173.

_______, David W. Gerbing and John E. Hunter (1987), On the Assessment of Unidimensional Measurement: Internal and External Consistency, and Overall Consistency Criteria. *Journal of Marketing Research*, 24 (November), 432-437.

_______ and David W. Gerbing (1988), Structural Equation Modeling in Practice: A Review and Recommended Two-Step Approach. *Psychological Bulletin*, 103 (May), 411-23.

Bagozzi, Richard P. (1980a), Performance and Satisfaction in an Industrial Sales Force: An Examination of their Antecedents and Simultaneity. *Journal of Marketing*, 44 (Spring), 65-77.

_______ (1980b), *Causal Models in Marketing*. New York: Wiley.

_______ (1981), An Examination of the Validity of Two Models of Attitude. *Multivariate Behavioral Research*, 16 (July), 323-359.

_______ and Lynn W. Phillips (1982), Representing and Testing Organizational Theories: A Holistic Construal. *Administrative Science Quarterly*, 27 (September), 459-489.

_______ (1984), A Prospectus for Theory Construction in Marketing. *Journal of Marketing*, 48 (Winter), 11-29

_______ and Hans Baumgartner (1994), The Evaluation of Structural Equation Models and Hypothesis Testing. In *Principles of Marketing Research*, R. Bagozzi ed., Cambridge MA: Blackwell, 386-422.

_______ and Todd F. Heatherton (1994), A General Approach to Representing Multifaceted Personality Constructs: Application to Self Esteem. *Structural Equation Modeling*, 1 (1), 35-67.

Bentler, Peter M. (1989), *EQS Structural Equations Program Manual*. Los Angeles: BMDP Statistical Software.

_______ (1990), Comparative Fit Indexes in Structural Models. *Psychological Bulletin*, 107 (March), 238-46.

Bohrnstedt, G. W. and T. M. Carter (1971), Robustness in Regression Analysis. In *Sociological Methodology*, H.L. Costner ed., San Francisco: Jossey-Bass 118-46.

Bollen, Kenneth A. (1989), *Structural Equations with Latent Variables*. New York: Wiley.

_______ and J. Scott Long (1993), *Testing Structural Equation Models*. Newbury Park, CA: SAGE Publications.

Browne, Michael W. (1982), Covariance Structures. In *Topics in Applied Multivariate Analysis*, Douglas M. Hawkins ed., Cambridge: Cambridge University Press.

_______ and Robert Cudeck (1993), Alternative Ways of Assessing Model Fit. In *Testing Structural Equation Models*, K. A. Bollen et al. eds, Newbury Park CA: SAGE Publications.

Burt, Ronald S. (1973), Confirmatory Factor-analysis Structures and the Theory Construction Process. *Sociological Methods and Research*, 2, 131-187.

_______ (1976), Interpretational Confounding of Unobserved Variables in Structural Equation Models. *Sociological Methods and Research*, 5 (August), 3-52.

Busemeyer, Jerome R. and Lawrence E. Jones (1983), Analysis of Multiplicative Combination Rules When the Causal Variables are Measured With Error. *Psychological Bulletin*, 93 (May), 549-62.

Calder, Bobby J. (1977), Focus Groups and the Nature of Qualitative Marketing Research. *Journal of Marketing Research*, XIV (August), 353-364.

Campbell, Donald T. and Donald W. Fiske (1959), Convergent and Discriminant Validation by the Multitrait-Multimethod Matrix. *Psychological Bulletin*, 56, 81-105.

_______ (1960), Recommendations for APA Test Standards Regarding Construct, Trait and Discriminant Validity. *American Psychologist*, 15, 546-553.

Cattell, R. B. (1973), *Personality and Mood by Questionnaire*. San Francisco: Jossey-Bass.

_______ (1978), *The Scientific use of Factor Analysis in Behavioral and Life Sciences*. New Your: Plenum.

Churchill, Gilbert A, Jr. (1979), A Paradigm for Developing Better Measures of Marketing Constructs. *Journal of Marketing Research*, 16 (February), 64-73.

_______ and J. Paul Peter (1984), Research Design Effects on the Reliability of Rating Scales: A Meta Analysis. *Journal* *of Marketing Research*, 21 (November), 360-75.

Cohen, Jacob and Patricia Cohen(1983), *Applied Multiple Regression/Correlation Analyses for the Behavioral Sciences*. Hillsdale, NJ: Lawrence Erlbaum.

Cote, Joseph A. and M. Ronald Buckley (1987), Estimating Trait, Method and Error Variance: Generalizing Across Seventy Construct Validation Studies. *Journal of Marketing Research*, 24 (August), 315-18.

_______ and M. Ronald Buckley (1988), Measurement Error and Theory Testing in Consumer Research: An Illustration of the Importance of Construct Validation. *Journal of Consumer Research*, 14 (March), 579-82.

Cronbach, Lee J. (1951), Coefficient Alpha and the Internal Structure of Tests. *Psychometrika*, 16 (September), 297-334.

_______, G. C. Gleser, H. Nada, and N. Rajaratnam (1972), *Dependability of Behavioral Measurements: Theory of Generalizability for Scores and Profiles*. New York: Wiley.

Darden, William R., S. Michael Carlson and Ronald D. Hampton (1984), Issues in Fitting Theoretical and Measurement Models in Marketing. *Journal of Business Research*, 12, 273-296.

DeVellis, Robert F. (1991), *Scale Development: Theory and Applications*. Newbury Park, CA: SAGE Publications.

Dillon, William R. (1986), Building Consumer Behavior Models with Lisrel: Issues in Applications. In *Perspectives on Methodology in Consumer Research*, D. Brinberg and R. J. Lutz Eds., New York: Springer-Verlag.

_______ and Matthew Goldstein (1984), *Multivariate Analysis: Methods and Applications*. New York: Wiley.

_______, Thomas J. Maden and Neil H. Firtle (1987), *Marketing Research in a Marketing Environment*. St. Louis: Times Mirror.

Dwyer, F. Robert and Sejo Oh (1987), Output Sector Munificence Effects on the Internal Political Economy of Marketing Channels. *Journal of Marketing Research*, 24 (November), 347-358.

Efron, B. (1981), Nonparametric Estimates of Standard Error: The Jackknife, the Bootstrap, and other Resampling Methods. *Biometrika*, 68, 589-599.

Fornell, Claes and David F. Larker (1981), Evaluating Structural Equation Models with Unobservable Variables and Measurement Error. *Journal of Marketing Research*, 18 (February), 39-50.

_______ and Fred L. Bookstein (1982), Two Structural Equation Models: LISREL and PLS Applied to Exit-Voice Theory. *Journal of Marketing Research*, 19 (November), 440-452.

Gerbing, David W. and James C. Anderson (1984), On the Meaning of Within-Factor Correlated Measurement Errors. *Journal of Consumer Research*, 11 (June), 572-580.

_______ and James C. Anderson (1988), An Updated Paradigm for Scale Development Incorporating Unidimensionality and Its Assessment. *Journal of Marketing Research*, 25 (May), 186-92.

_______, Janet G. Hamilton and Elizabeth B. Freeman (1994), A Large-scale Second-order Structural Equation Model of the Influence of Management Participation on Organizational Planning Benefits. *Journal of Management*, 20, 859-885.

Green, Samuel B, Robert W. Lissitz and Stanley A. Mulaik (1977), Limitations of Coefficient Alpha as an Index of Test Unidimensionality. *Educational and Psychological Measurement*, 37 (October), 827-38.

Hattie, J. (1985), Methodology Review: Assessing Unidimensionality of Tests and Items. *Applied Psychological Measurement*, 9, 139-164.

Heeler, Roger M. and Michael L. Ray (1972), Measure Validation in Marketing. *Journal of Marketing Research*, 9 (November), 361-70.

Heise, D. R. and G. W. Bohrnstedt (1970), Validity, Invalidity and Reliability. In *Sociological Methodology*, E. F. Borgatta and G. W. Bohrnstedt eds., New York: Jossey Bass.

Herche, Joel and Brian Engelland (1996), Reversed-Polarity Items and Scale Unidimensionality. *Journal of the Academy of Marketing Science*, 24 (Fall), 366-74.

Hoelter, J. W. (1983), The Analysis of Covariant Structures: Goodness of Fit Indices. *Sociological Methods and Research*, 11, 325-344.

Hunter, John Edward (1973), Methods of Reordering the Correlation Matrix to Facilitate Visual Inspection and Preliminary Cluster Analysis. *Journal of Educational Measurement*, 10 (Spring), 51-61.

_______ and David W. Gerbing (1982), Unidimensional Measurement, Second-Order Factor Analysis and Causal Models. In *Research in Organizational Behavior*, Vol. IV, Barry M. Staw and L. L. Cummings eds., Greenwich CT: JAI Press, 267-320.

Jaccard, James, Robert Turrisi and Choi K. Wan (1990), *Interaction Effects in Multiple Regression*. Newbury Park, CA: SAGE Publications.

_______ and C. K. Wan (1995), Measurement Error in the Analysis of Interaction Effects Between Continuous Predictors Using Multiple Regression: Multiple Indicator and Structural Equation Approaches. *Psychological*

Bulletin, 117 (2), 348-357.

Jöreskog, Karl G. (1970), A General Method for Analysis of Covariance Structures. *Biometrika*, 57, 239-251.

_______ (1971) Simultaneous Factor Analysis in Several Populations. *Psychometrika*, 57, 409-426.

_______ (1993), Testing Structural Equation Models. In *Testing Structural Equation Models*, Kenneth A. Bollen and J. Scott Long eds., Newbury Park, CA: SAGE Publications.

_______ and Dag Sörbom (1996a), *Prelis 2 User's Reference Guide*. Chicago: Scientific Software International, Inc.

_______ and Dag Sörbom (1996b), *Lisrel 8 User's Reference Guide*. Chicago: Scientific Software International, Inc.

Kenny, David (1979), *Correlation and Causality*. New York: Wiley.

Kumar, Ajith and William R. Dillon (1987a), The Interaction of Measurement and Structure in Simultaneous Equation Models with Unobservable Variables. *Journal of Marketing Research*, 24 (February), 98-105.

_______ and William R. Dillon (1987b), Some Further Remarks on Measurement-Structure Interaction and the Unidimensionality of Constructs. *Journal of Marketing Research*, 24 (November), 438-44.

Lord, F. M. (1980), *Applications of Item Response Theory to Practical Testing Problems*. New York: Erlbaum.

_______ and M. R. Novick (1968), *Statistical Theories of Mental Test Scores*. Reading, MA: Addison Wesley.

McClelland, G. H. and C. M. Judd (1993), Statistical Difficulties of Detecting Interactions and Moderator Effects. *Psychological Bulletin*, 114 (2), 376-390.

McDonald, R. P. (1981), The Dimensionality of Tests and Items. *British Journal of Mathematical and Statistical Psychology*, 34, 100-117.

Neter, John, William Wasserman and Michael H. Kunter (1985), *Applied Linear Statistical Models*. Homewood, IL: Irwin.

Novick, M. R. and C. Lewis (1967), Coefficient Alpha and the Reliability of Composite Measurements. *Psychometrika*, 32, 1-13.

Nunnally, Jum C. (1978), *Psychometric Theory*, 2nd Ed. New York: McGraw-Hill.

Peter, J. Paul (1979), Reliability: A Review of Psychometric Basics and Recent Marketing Practices. *Journal of Marketing Research*, 16 (February), 6-17.

_______ (1981), Construct Validity: A Review of Basic Issues and Marketing Practices. *Journal of Marketing Research*, 18 (May), 133-45.

_______ and Gilbert A. Churchill, Jr. (1986), Relationships Among Research Designs Choices and Psychometric Properties of Rating Scales: A Meta Analysis. *Journal of Marketing Research*, 23 (February), 1-10.

Ping, R. (1998a), Some Suggestions for Validating Measures Involving Unobserved Variables and Survey Data. *1998 Winter American Marketing Association Educators' Conference*, Chicago: American Marketing Association.

_______ (1998b), Improving the Detection of Interactions in Selling and Sales Management Research. *Journal of Personal Selling and Sales Research*, XVI (Winter), 53-64.

Ray, John J. (1983), Reviving the Problem of Acquiescent Response Bias. *Journal of Social Psychology*, 121, 81-96.

Rindskopf, David and Tedd Rose (1988), Some Theory and Applications of Confirmatory Second-order Factor Analysis. *Multivariate Behavioral Research*, 23 (January), 51-67.

Rock, D. A., C. E. Werts, R. L. Linn and K. G. Jöreskog (1977), A Maximum Likelihood Solution to the Errors in Variables and Errors in Equations Model. *The Journal of Multivariate Behavioral Research*, 12 (April), 187-197.

Rusbult, Carl E., Dan Farrell, Glen Rogers and Arch G. Mainous III (1988), Impact of Exchange Variables on Exit, Voice, Loyalty, and Neglect: An Integrative Model of Responses to Declining Job Satisfaction. *Academy of Management Journal*, 31 (September), 599-627.

Saris, W. E., W. M. de Pijper and P. Zegwaart (1978), Detection of Specification Errors in Linear Structural Equation Models. In *Sociological Methodology*, K. E. Schuesster, ed., San Francisco: Jossey-Bass.

Smith, Kent W. (1974), On Estimating the Reliability of Composite Indexes Through Factor Analysis. *Sociological Methods and Research*, 2 (May), 485-510.

Sörbom, D. (1975), Detection of Correlated Errors in Longitudinal Data. *British Journal of Mathematical and Statistical Psychology*, 28, 138-51.

Steiger, J.H. (1990), Structural Model Evaluation and Modification: An Interval Estimation Approach. *Multivariate Behavioral Research*, 25, 173-180.

Tyron, R. C. (1935), A Theory of Psychological Components- An Alternative to Mathematical Factors. *Psychological Review*, 42, 425-454.

Warren, R. D., J. K. White and W. A. Fuller (1974), An Errors-in-Variables Analysis of Managerial Role Performance. *Journal of the American Statistical Association*, 69, 886-893.

Werts, C.E., R.L. Linn and K.G. Jöreskog (1974), Intraclass Reliability Estimates: Testing Structural Assumptions. *Educational and Psychological Measurement*, 34, 25-33.

Williams, Larry J. and John T. Hazer (1986), Antecedents and Consequences of Satisfaction and Commitment in Turnover Models: A Reanalysis Using Latent Variable Structural Equation Methods. *Journal of Applied Psychology*, 71 (May), 219-231.

APPENDIX A- Consistency Improvement using Summed First Derivatives

A measure of the latent variable X with eight items was used in a marketing survey that produced more than 200 usable responses. The first derivatives with respect to the error terms (Var(e)' s in Equation 1)from a single construct measurement model of X, and their sum without regard to sign for each item, are shown in Table A1. The item with the largest Table A1 column sum without regard to sign (x_4) was omitted, and the measurement model was re estimated to produce the Table A2 first derivatives. This process was repeated until RMSEA was .08 or less (see Table A4). An investigation of all other measurement models with of five items (not shown) produced combinations of items that were less consistent (i.e., they had worse model-to-data fit statistics), suggesting the Table A4 items were maximally consistent.

However, maximizing consistency does not necessarily maximize reliability or Average Variance Extracted (AVE). The items with maximum reliability and AVE were x_4, x_5, x_6, x_7, and x_8 (Reliability = .884 and AVE = .606, but $\chi^2 = 25$, df = 5, p-value = .0001, RMSEA = .135).

There is no guidance for trading off reliability and consistency. In the present case the reliabilities of both the Table A4 itemization and x_4, x_5, x_6, x_7, and x_8 would likely be judged acceptable. However AVE for the Table A4 itemization is only slightly above the suggested cutoff (i.e., .5), and x_4 through x_8 are marginally consistent. In cases where reliability and consistency diverge, I would suggest using the higher reliability itemization first.

Table A1- First Derivatives for the Eight Item Measure

	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8
x_1	0.000	-0.439	-0.025	-0.086	0.047	0.006	0.010	0.371
x_2	-0.439	0.000	-0.272	0.287	0.217	0.042	-0.200	0.143
x_3	-0.025	-0.272	0.000	-0.527	0.184	0.364	0.422	-0.207
x_4	-0.086	0.287	-0.527	0.000	-0.943	0.505	0.534	0.144
x_5	0.047	0.217	0.184	-0.943	0.000	0.222	0.359	0.019
x_6	0.006	0.042	0.364	0.505	0.222	0.000	-0.929	-0.187
x_7	0.010	-0.200	0.422	0.534	0.359	-0.929	0.000	-0.113
x_8	0.371	0.143	-0.207	0.144	0.019	-0.187	-0.113	0.000
Sum[a]	0.983	1.600	2.000	3.027	1.991	2.254	2.565	1.184

$\chi^2 = 86$ df = 20 p-value = 0 RMSEA[b] = .123 Reliability = .860 AVE = .442

APPENDIX A- Consistency Improvement using Summed First Derivatives (Continued)

Table A2- First Derivatives with x_4 Deleted

	x_1	x_2	x_3	x_5	x_6	x_7	x_8
x_1	0.000	-0.442	-0.064	-0.057	0.037	0.044	0.354
x_2	-0.442	0.000	-0.287	0.129	0.214	-0.067	0.195
x_3	-0.064	-0.287	0.000	-0.172	0.319	0.382	-0.313
x_5	-0.057	0.129	-0.172	0.000	0.090	0.231	-0.252
x_6	0.037	0.214	0.319	0.090	0.000	-0.544	0.012
x_7	0.044	-0.067	0.382	0.231	-0.544	0.000	0.112
x_8	0.354	0.195	-0.313	-0.252	0.012	0.112	0.000
Sum[a]	0.998	1.334	1.537	0.933	1.217	1.381	1.239

χ^2 = 56 df = 14 p-value = .44E-6 RMSEA[b] = .117 Reliability = .828 AVE = .416

Table A3- First Derivatives with x_3 and x_4 Deleted

	x_1	x_2	x_5	x_6	x_7	x_8
x_1	0.000	-0.445	-0.086	0.045	0.054	0.304
x_2	-0.445	0.000	0.036	0.190	-0.103	0.107
x_5	-0.086	0.036	0.000	0.114	0.270	-0.383
x_6	0.045	0.190	0.114	0.000	-0.252	-0.013
x_7	0.054	-0.103	0.270	-0.252	0.000	0.096
x_8	0.304	0.107	-0.383	-0.013	0.096	0.000
Sum[a]	0.937	0.883	0.891	0.616	0.776	0.904

χ^2 = 36 df = 9 p-value = .38E-4 RMSEA[b] = .116 Reliability = .814 AVE = .433

Table A4- First Derivatives with x_1, x_3 and x_4 Deleted

	x_2	x_5	x_6	x_7	x_8
x_2	0.000	-0.026	0.110	-0.180	0.079
x_5	-0.026	0.000	0.104	0.252	-0.352
x_6	0.110	0.104	0.000	-0.233	0.064
x_7	-0.180	0.252	-0.233	0.000	0.173
x_8	0.079	-0.352	0.064	0.173	0.000

χ^2 = 5.89 df = 5 p-value = .136 RMSEA[b] = .028 Reliability = .835 AVE = .509

[a] Without regard to sign

[b] .05 suggests close model-to-data fit, .051-.08 suggests acceptable model-to-data fit (Brown and Cudeck 1993, Jöreskog 1993).

APPENDIX B- Ordered Similarity Coefficients and Consistency

Similarity coefficients for the eight items analyzed in Appendix A are shown in Table B, in descending summed similarity. For example, x_4 has a similarity of 1.00 with itself, and .97 with x_5, etc. It is the most similar to all the items (i.e., while the sums of similarity coefficients are not shown, it is obvious that x_4 has the largest summed similarity), and is most similar to x_5. x_5 is the next most similar item and after x_4 it is most similar to and x_6.

However, ordered similarity coefficients do not necessarily suggest maximally consistent item clusters. The most similar items are x_4, x_5, x_6, x_7, and x_8, (each have .9 or higher coefficients with the others). Item x_3 is less similar with .91 or lower coefficients, and x_2 and x_1 have one or more coefficients below the suggested .8 cutoff. Nevertheless the five items with the highest consistency were x_2, x_5, x_6, x_7, and x_8 (see Appendix A).

Table B-Ordered Similarity Coefficients for the Appendix A Items[a]

	x_4	x_5	x_6	x_7	x_8	x_3	x_2	x_1
x_4	100	97	94	93	93	91	80	71
x_5	97	100	94	93	93	87	79	69
x_6	94	94	100	97	93	84	80	69
x_7	93	93	97	100	92	83	82	69
x_8	93	93	93	92	100	88	76	60
x_3	91	87	84	83	88	100	81	66
x_2	80	79	80	82	76	81	100	81
x_1	71	69	69	69	60	66	81	100

[a] Table entries are coefficients times 100

APPENDIX C- An Example Scenario

Scenario analysis is composed of instructions (see Exhibit C), a scenario (titled "Research Material" in Exhibit C), a questionnaire containing measures for the study constructs which is attached to the instructions (not shown), and student subjects. The scenario manipulates the independent variables, and the questionnaire measures the manipulations and the dependent variables.

In the Exhibit C scenario each student received the Instructions/Research Materials sheet with the questionnaire attached. The Research Material shown in Exhibit C has been truncated at the ellipses to conserve space, and each student received a Research Material section showing only one of the two possible choices in each parenthesis. The Exhibit C experiment had 8 treatments (see the last paragraph of the Research Material), each with two levels (represented by the alternatives in parentheses), so there were 256 (= 2^8) different Research Materials, one for each treatment group. Ideally treatment groups should be homogenous within, and there should be more than one subject per treatment. However, this particular scenario involved one subject per treatment and nonhomogeneous student subjects (see the results in Appendix D).

Exhibit C- A Scenario

INSTRUCTIONS: Please read the following material, and then respond to the statements that follow it. Your responses are anonymous and very important to the development of a study of personal selling.

RESEARCH MATERIAL

Please attempt to place yourself in the position of X, the major character in the following short story. Try to imagine that person's feelings and attitudes as vividly as you can, considering what it would be like to be in their situation. You may need to read the story several times before you are completely familiar with the details of the situation. Then respond to the statements that follow the story, indicating how you would react if you were in that situation. There are no "right" or "wrong" answers. It is your own, honest opinion of how X would feel and act that we want.

Imagine that you are X. You are working for a financial services company. The company sells mutual funds and other investments. It helps clients manage their personal and family assets using offices located around the country. Clients seek the company's advice and investment products to maintain and build their net worth for retirement, college for their children, etc.

You are an account representative for, among other things, the company's mutual fund products that include stock and bond funds, and funds made up of securities from foreign companies. You are very good at advising clients regarding their financial planning. You and the company have (a common, different) goal-- (satisfied customers, you want satisfied customers and they want brokerage fees). You are also paid a very (attractive, *un*attractive) combination of salary and commissions that (generously, does *not*) compensate(s) you for all the preparation and work that you do for the company. The company's policies and procedures regarding performance evaluation and feedback, promotion, vacation, health care, etc. are (very, *not*) fair compared to other companies. These policies and procedures are administered very (fairly, *un*fairly): you see (no) favoritism in promotions, for example, (and, or) inconsistent administration of these policies and procedures (any-, every-)where. You are treated with (great, *no*) respect (and, or) concern for your feelings by company management.

You have worked for the company for (seven years, three months) now, and have devoted many of these years to developing your client base. (You have spent many nights and weekends, Some of this time has been devoted to)

APPENDIX C- An Example Scenario (Continued)

learning the company's products and services, and how to serve clients with these products and services, (that could have been spent having fun; Some of this time has been spent developing your client base).

.

.

.

Things at work had been fine, but in the past week a problem developed. Your manager called you to say that you will be asked to give several of your best clients to the newly hired account representatives. They currently go too long without commissions. In addition, you will be asked to help train these new account representatives. This would reduce the available time you have to find replacement clients, and reduce your ability to serve your existing clients.

Remember, you have worked for this company a (*long*, *short*) time. You and the company have (the *same*, very *different*) goals. Your compensation is very (*fair*, *un*fair). The company's policies and procedures are very (fair, *un*fair). These policies and procedures are administered very (fairly, *un*fairly). You are treated with (great, *no*) respect (and, or) concern for your feelings by your company's management. Other potential employers are very (attractive, *un*attractive). Changing jobs would require *(a lot* of, *little*) effort (and, or) risk.

APPENDIX D- Scenario Analysis Results Comparison

The Appendix C scenario was administered to more than 200 students, and its questionnaire was also mailed to a sample from a non student population, and this generated more than 200 responses. A psychometric comparison of the scenario analysis results and the survey data results using the same questionnaire is shown in Table D. They are similar enough to suggest that scenario analysis may be useful for measure debugging, and preliminary model evaluation.

Table D- Comparison of Scenario and Survey Data from a Common Questionnaire Using Factor Analysis

Scenario Data:

FACTOR	1	2	3	4	5
EX6	.858				
EX4	.851				
EX2	.839				
EX7	.839				
EX5	.826				
EX1	.770				
EX8	.769				
EX3	.730				
IN8		.933			
IN3		.897			
IN5		.897			
IN6		.887			
IN1		.869			
IN4		.861			
IN7		.683			
IN2		.601			
AL5			.820		
AL6			.768		
AL3			.743		
AL2			.739		
AL4			.732		
AL7			.729		
AL1			.701		
SA7			.814		
SA3				.780	
SA2				.771	
SA8				.750	
SA6				.721	
SA4				.718	
SA1				.657	
SA5				.518	
SC2					.823
SC4					.778
SC5					.721
SC6	-.406				.711
SC1					.692
SC3	-.443				.642
Eigen-value	13.24	5.93	3.10	2.35	2.06
Pct. Var	35.8	16.0	8.4	6.4	5.6

Field Survey Data:

	1	2	3	4	5
EX7	.841				
EX3	.829				
EX2	.821				
EX4	.815				
EX5	.814				
EX1	.807				
EX6	.778				
EX8	.771				
SA8		.850			
SA7		.848			
SA4		.809			
SA6		.794			
SA3		.747			
SA2		.746			
SA1		.703			
SA5		.675			
IN5			.906		
IN8			.901		
IN3			.879		
IN4			.876		
IN6			.873		
IN1			.823		
IN7			.680		
IN2			.646		
AL5				.778	
AL1				.771	
AL3				.768	
AL2				.761	
AL7				.759	
AL4	-.415			.752	
AL6				.646	
SC5					.797
SC4					.784
SC6					.768
SC3					.743
SC2					.637
SC1					.635
	16.23	5.87	2.79	1.85	1.79
	43.9	15.9	7.6	5.0	4.9

APPENDIX E- Average Variance Extracted

A marketing survey involving the latent variables T, U, V, W, and the interaction UxT produced more than 200 usable responses. After item omissions to attain sufficient consistency, the measures for these latent variables were judged to be unidimensional, valid and reliable. A second marketing survey involving the latent variables A, B, C, D, and E also produced more than 200 usable responses. The measure for A was an established measure, and omitting items in several other measures to attain acceptable consistency was judged to degrade their content or face validity. Thus, these measures were used with no item omissions. The unidimensionality of A, B, C, D, and E was gauged using maximum likelihood exploratory common factor analysis. Each of the measures for A through E produced one factor with an eigenvalue greater than one, which suggested their unidimensionality. Table E presents the reliabilities and average extracted variance estimates for these variables. Since the model-to-data fit of the structural equation model for A through E was below acceptability, both the average variance extracted and the latent variable reliability (LV Reliability) estimates are approximations.

Table E- Reliability and AVE Comparisons

Measure	U	T	UxT	V	W	A	B	C	D	E
LV Reliability[a]	.946	.635	.686	.817	.928	.933	.926	.927	.949	.962
SPSS Reliability[b]	.942	.609	.749[e]	.818	.925	.941	.929	.917	.947	.968
SEA AVE[c]	.781	.384	.136	.534	.765	.672	.692	.671	.792	.721
Eigenvalue AVE[d]	.737	.341	.166	.475	.731	.643	.637	.588	.760	.710

[a] Using Equation 4.

[b] Reliabilities using raw data and SPSS, except for the reliability of UxT.

[c] Using Equation 6.

[d] AVE estimated using the Footnote 6 approach involving the percent explained variance of each measure from Maximum Likelihood exploratory common factor analyses (except for UxT-- see Footnote e below).

[e] Using Equation 7.

APPENDIX F- A Structural Equation Model with Single Indicators

The data from the second marketing survey described in Appendix E was used to estimate the variables A through D's associations with E using single indicator structural equation analysis (SEA). A single averaged indicator was used for each of the variables A, B, C, D, and E. To use this single indicator approach, the indicators for each latent variable X should be unidimensional using maximum likelihood exploratory common factor analysis and criteria such only one factor with an eigenvalue greater than one. Next the indicators for X, x_1, x_2, ... , x_n, should be averaged, and the value $(x_1+x_2+ ... +x_n)/n$ should be added to each case. Then the variance of X, Var(X), and the reliability of X, α_X, should be determined using SAS, SPSS, etc. Next the loading of each x_i on X should be determined using maximum likelihood exploratory common factor analysis (i.e., using SAS, SPSS, etc.), each loading should be scaled (see below), and the loadings should be averaged to form Λ_X. Finally, the averaged indicator should be specified in the structural equation analysis model (i.e., the measurement or structural model) with a fixed loading equal to Λ_X and a fixed measurement error equal to Var(X)(1-α).

The above steps were taken for the variables A through E, and the results are shown in Table F1. For emphasis, the exploratory common factor analysis used maximum likelihood extraction, and the LISREL 8 estimation used maximum likelihood estimation. In addition before they were averaged, the maximum likelihood exploratory common factor analysis loadings were scaled by dividing each measure's loadings by its measure's maximum loading (i.e., each of A's loadings, l_a, was replaced by $l_a/\max(l_a)$, where $\max(l_a)$ is the largest loading on A); each of B's loadings, l_b, was replaced by $l_b/\max(l_b)$, where $\max(l_b)$ is the largest loading on B; etc.), which is required to syncronize the variances of indicators with that of their construct (this is similar to fixing an indicator at one to provide a metric [for non-unit-variance] A through E in structural equation analysis).

Table F2 shows the results of estimating the same model with A through E each specified using their multiple indicators. Since the reliabilities of A through E were above .9, Table F3 shows the results of replacing Λ with $(\alpha)^{1/2}$ in the Table F1 model. To obtain these results, the coefficient alphas were determined for A through E using SPSS, and the square roots of these values replaced the Λ's in the Table F1 model. (Note: A through E were unstandardized--their variances were not 1.)

APPENDIX F- A Structural Equation Model with Single Indicators (Continued)

Table F1- Single Indicator Coefficient Estimates Using Exploratory Common Factor Analysis

Dependent Variable= E

	A	B	C	D	χ^2/df	GFI[b]	AGFI[b]	CFI[c]	RMSEA[d]
b_i[a]	-0.478	0.358	0.000	0.061	0/0	-------(not applicable)-------			
t-value	-6.94	5.18	0.00	0.87					

Table F2- Multiple Indicator Coefficient Estimates

Dependent Variable= E

	A	B	C	D	χ^2/df	GFI[b]	AGFI[b]	CFI[c]	RMSEA[d]
b_i[a]	-0.465	0.357	-0.025	0.072	1079/517	.774	.740	.928	.070
t-value	-6.90	5.30	-0.38	1.11					

Table F3- Single Indicator Coefficient Estimates Using $(\alpha)^{1/2}$

Dependent Variable= E

	A	B	C	D	χ^2/df	GFI[b]	AGFI[b]	CFI[c]	RMSEA[d]
b_i[a]	-0.478	0.358	0.000	0.061	0/0	-------(not applicable)-------			
t-value	-6.94	5.18	0.00	0.87					

[a] Standardized estimates.

[b] Shown for completeness only-- GFI and AGFI may be inadequate for model-to-data fit assessment in larger models (see Anderson and Gerbing 1984).

[c] .90 or better indicates acceptable model-to-data fit (see McClelland and Judd 1993).

[d] .05 suggests close model-to-data fit, .051-.08 suggests acceptable model-to-data fit (Brown and Cudeck 1993, Jöreskog 1993).

Revision History:

2010: Added "Notes" to update paper for recent results--search on "Notes" to locate.

2006: Paper was written in 1999, edited to reduce length, accepted in 2000, and published in 2004. This unreduced version was selectively revised to make it current through 2006.

BUT WHAT ABOUT CATEGORICAL (NOMINAL) VARIABLES IN LATENT VARIABLE MODELS?

ABSTRACT

The paper suggests an approach for specification, estimation and interpretation of a categorical or nominal exogenous (independent) variable in theory or hypothesis tests of latent variable models with survey data. An example using survey data is provided.

Anecdotally, categorical variables (e.g., Marital Status) are ubiquitous in applied marketing research. However, they are absent from published theory (hypothesis) tests of latent variable models using survey data.

One plausible explanation is there is no explicit provision for "truly" categorical variables in the popular structural equation (covariant structure) analysis software packages (e.g., LISREL, EQS, AMOS, etc.). There, the term "categorical variable" means an ordinal variable (e.g., an attitude measured by a Likert scale) (e.g., Jöreskog and Sörbom 1996), rather than a nominal variable such as Marital Status.

Further, normality is a fundamental assumption in covariance structural analysis (e.g., in LISREL, EQS, Amos, etc.) for maximum likelihood estimation, the preferred estimator for hypothesis tests involving latent variables and survey data (e.g., Jöreskog and Sörbom 1996). A (truly) categorical independent variable is typically estimated using "dummy" variables that are not normally distributed. (For example, while case values for the categorical variable Marital Status, for example, might be 1 for Single, 2 for Married, 3 for Divorced, etc., new variables, for example Dummy_Single, Dummy_Married, etc., are created and estimated instead of Marital Status. Dummy_Single might have a case value of 1 if Marital Status = Single, and 0 otherwise, Dummy_Married might have cases that have the value 1 if Marital Status = Married, and 0 otherwise, etc.)

There are other less obvious barriers in survey-data theory tests to adding (truly) categorical exogenous variables to models that also contain latent variables, including determining the significance of a categorical variable when its dummy variables are estimated instead. If each dummy variable is significant (or nonsignificant), it seems reasonable to conclude that the categorical variable from which the dummies were created is significant (or nonsignificant). However, if some dummy variables are significant but some are not, there is no guidance for estimating the significance of the categorical variable from which they were created. In addition, interpreting a significant categorical variable can involve interpreting changes in intercepts, parameters that are not usually estimated in theoretical model testing.

Several approaches have been suggested for estimating categorical variables (e.g., dummy variable regression, logistic regression, latent category models, etc.). However, there is no guidance for estimating a "mixed" model--one that combines a categorical exogenous variable with latent variables--in theory (hypothesis) tests involving survey data.

This paper addresses these matters. It suggests a specification for a categorical variable in theory (hypothesis) tests of latent variable models involving survey data. It also discusses the estimation and interpretation of a categorical variable in these "mixed" models.

AN EXAMPLE

To expedite the presentation of these matters, we will use a real-world data set involving buyer-seller relationship Satisfaction (SAT), Alternative relationship attractiveness (ALT), and Exit propensity (EXI). Data (200+ cases) was collected in a survey to test a larger model in which Satisfaction and Alternative Unattractiveness were hypothesized to be negatively associated with Exiting. SAT, ALT and EXI were measured using multiple item measures (Likert scales). The resulting latent variables, SAT, ALT and EXI were judged to be valid and reliable, and the itemizations of each were judged to be internally and externally consistent in the Anderson and Gerbing (1988) sense.[1]

The structural equation

$$EXI = b_1SAT + b_2ALT + \zeta, \qquad (1$$

where b_i are structural coefficients and ζ is structural disturbance, was estimated using LISREL and maximum likelihood estimation, and SAT and ALT were significantly (negatively) associated with EXI as shown in Table A.

SAT was measured with five-point Likert-scaled items that each could be analyzed as a categorical variable. (E.g., the SAT indicator Sa2 had 5 categories: those respondents who strongly agreed they were satisfied, those who agreed they were satisfied, those who were neutral, etc.) For pedagogical purposes SAT was replaced in Equation 1 with one of its high loading indicators, Sa2. The resulting model was estimated, and Sa2 and ALT were significantly and negatively associated with EXI, also as shown in Table A.

Next, dummy variables for the categories of Sa2 (i.e., category 1 = strongly dissatisfied, category 2 = dissatisfied, category 3 = neutral, category 4 = satisfied, and category 5 = very satisfied) were created. Specifically, in each case, Sat_Dummyi = 1 if Sa2 = i (i = 1 to 5) in that case. Otherwise, Sat_Dummyi = 0. Thus for example, in each

[1] Other study details are omitted to skirt matters such as conceptual and operational definitions, etc. that were judged to be of minimal importance to the present purposes.

case where Sa2 = 1, Sat_Dummy1 was assigned the value of 1. For those cases where Sa2 equaled some other value (i.e., 2, 3, 4 or 5), Sat_Dummy1 was assigned 0.

Equation 1 was altered to produce the structural model

$$\begin{aligned} EXI = {} & b_{11}Sat\_Dummy1 + b_{12}Sat\_Dummy2 + b_{13}Sat\_Dummy3 \\ & + b_{14}Sat\_Dummy4 + b_{15}Sat\_Dummy5 + b_2'ALT + \zeta', \qquad (2 \end{aligned}$$

where b_{1j} and b_2' (j = 1 to 5) are structural coefficients, and ζ' is structural disturbance.

Unfortunately, Equation 2 currently cannot be estimated satisfactorily using LISREL (or other popular covariant structure packages such as EQS, AMOS, etc.). The covariance matrix produced by the dummy variables is not positive definite.

The usual "remedy" is to estimate Equation 2 with one dummy variable omitted (e.g., Blalock 1979). However, this approach is unsatisfactory for theory testing because omitting a dummy variable in Equation 2 alters the significances of the remaining dummy variables, depending on which dummy variable is omitted. For example, see the significances in Tables B and C of Sat_Dummy2 or Sat_Dummy4 when Sat_Dummy1 or SAT_Dummy5 was omitted from Equation 2.

Ping (1996) proposed a latent variable estimation approach that will estimate Equation 2 without omitting dummy variables. The approach, Latent Variable Regression, adjusts the Equation 2 variance-covariance matrix for the measurement errors in ALT and EXI using Equation 2 measurement model loadings and measurement error variances. The resulting error-disattenuated variance-covariance matrix is then input to OLS regression. This approach was judged to be acceptably unbiased and consistent in the Ping (1996) article.

In order to use Latent Variable Regression to estimate Equation 2, the error-adjusted covariance matrix for Equation 2, and "regression through the origin" was used (to accommodate the collinearity of the dummy variables--see Blalock 1979). Specifically, the (error attenuated) variances and covariances of ALT, EXI and the five indicators for the SAT dummy variables were obtained using SPSS. Next, the measurement model for Equation 2 was estimated using the (consistent) indicators for ALT and EXI, and single indicators for the five SAT dummy variables (i.e., Sat_Dummy1, Sat_Dummy2, etc.), with LISREL and maximum likelihood estimation. Then, the loadings and measurement error variances from this measurement model were used to adjust the SPSS variances and covariances of ALT, EXI and the SAT dummy variables using equations proposed by Ping (1996) such as

$$\mathrm{Var}(\xi_X) = (\mathrm{Var}(X) - \theta_X)/\Lambda_X^2$$

and

$$\mathrm{Cov}(\xi_X,\xi_Z) = \mathrm{Cov}(X,Z)/\Lambda_X\Lambda_Z ,$$

where $\mathrm{Var}(\xi_X)$ is the error adjusted variance of X, Var(X) is the error attenuated variance of X (available from SAS, SPSS, etc.), $\Lambda_X = \mathrm{avg}(\lambda_{X1} + \lambda_{X2} + ... + \lambda_{Xn})$, avg = average, and $\mathrm{avg}(\theta_X = \mathrm{Var}(\varepsilon_{X1}) + \mathrm{Var}(\varepsilon_{X2}) + ... + \mathrm{Var}(\varepsilon_{Xn}))$, ($\lambda$'s and ε_X's are the measurement model loadings and measurement error variances--1 and 0 respectively--for the SAT dummy variables, and n = the number of indicators of the latent variable X), $\mathrm{Cov}(\xi_X,\xi_Z)$ is the error adjusted covariance of X and Z, and Cov(X,Z) is the error-attenuated covariance of X and Z.[2]

[2] These equations make the classical factor analysis assumptions that the measurement errors are independent of each other, and the x_i's are independent of the measurement errors. The indicators for X and Z must be consistent in the Anderson and Gerbing (1988) sense.

The resulting error-adjusted variance-covariance matrix for Equation 2 was then input to SPSS' OLS regression procedure, and the results are shown in Table D.

DISCUSSION

The dummy variables for categories 4 and 5 (Sat_Dummy4 and Sat_Dummy5) in Table D were nonsignificant, while the other dummy variables were significant. Comparing the Table D results for ALT to those from Tables B and C, the statistics for the coefficient of ALT were practically unaffected by omitting a single dummy variable, or by using regression through the origin. Also note that the unstandardized coefficient for Sat_Dummy2 with Sat_dummy1 omitted in Table B was Sat_Dummy2 - Sat_Dummy1 (within rounding), in Table D. Similarly, the unstandardized coefficient for Sat_Dummy3 with Sat_dummy1 omitted in Table B was Sat_Dummy3 - Sat_Dummy1, in Table D within rounding. Similarly, the other Table B dummy variables were the difference within rounding between their Table D value and Sat_Dummy1. For this reason Sat_Dummy1 is sometimes referred to as a "reference variable." In Table C Sat_Dummy5 is the reference variable for the unstandardized coefficient values shown there.

However, the interpretation or "meaning" of the unstandardized coefficients of the dummy variables is slightly different from the unstandardized coefficient of ALT. For example, the signs on the Table D unstandardized coefficients of the dummy variables have no meaning. Specifically, Sat_Dummy1, for example, is not "positively" associated with EXI. The unstandardized coefficient of Sat_Dummy1 is the absolute value of the change in the mean of EXI "caused" (associated) with Sat_Dummy1 (1.51). In this case, the mean of EXI changed in absolute value for the "very dissatisfied" category by the amount 1.51. Similarly, the absolute value of the change in EXI for the "very satisfied" (Sat_Dummy5) was .15. (Nevertheless, the "direction" of the associations of the set of dummy variables with EXI can be inferred--see below).

THE SATISFACTION-EXITING HYPOTHESIS

Satisfaction was hypothesized to be associated with Exiting, but so far we have estimated only variables derived from Satisfaction, its dummy variables, and several of them were nonsignificant. To estimate the effect of Satisfaction on Exiting using the dummy variables, the coefficients of the dummies were aggregated using a weighted average,[3] and the results are shown in Table E. Interpreting this aggregated result, Satisfaction, estimated as a categorical variable, was significantly associated with Exiting as hypothesized.

The "direction" of this association can be inferred by linearly ordering the unstandardized coefficients of the dummy variables from low to high. In this case EXI (i.e., its means in absolute value) decreased as the category represented by the dummy variables increased, which is consistent with the Table A result that Satisfaction is negatively associated with Exiting. Specifically, in the lower satisfaction categories, Sat_Dummy1 and Sat_Dummy2, Exiting was higher (i.e., the means were higher), while in the higher satisfaction categories, Sat_Dummy4 and Sat_Dummy5, Exiting was lower.

COMMENTS

[3] An overall F test of the "effect" of the dummies (e.g., $F = [(R_2^2 - R_1^2)/(k_2 - k_1)] / [(1- R_2^2)/(N - k_2 - 1)]$ where Ri^2 is R Square (or Squared Multiple Correlation), i = 1 denotes the model with the dummies omitted, i = 2 denotes the model with the dummies included, k_i is the number of exogenous variables (predictors), and N is the number of cases--see for example Jaccard, Turrisi and Wan, 1990) is inappropriate because R^2's are not comparable between intercept and no intercept models (Hahn 1977). (In addition, R^2 is usually incorrectly calculated in no intercept models--see Gordon 1981).

Anecdotally, it is believed that regression through the origin (no intercept regression) is biased, perhaps because its R^2 appears to be biased (Gordon 1981)--especially when the intercept is likely to be non-zero. However, experience suggests that with dummy variables, coefficient estimates are consistent between intercept regression and no intercept regression. For example, the coefficient estimates from omitting a dummy variable (which used intercept regression) such as those in Tables B and C were the difference between the omitted dummy variable's coefficient and the other coefficients in the no intercept results shown in Table D, within rounding). Also, the Equation 2 coefficients for ALT were practically unaffected by regression through origin (e.g., see Table B and Table D).

However, these results do not "prove" anything. They merely hint that the suggested approach may be useful for a categorical exogenous variable[4] in theory (hypothesis) tests of latent variable models with survey data. Nevertheless, as an additional example, ALT was estimated as a categorical variable with the results shown in Table F. These results paralleled those from estimating Satisfaction. For example, the Equation 2 coefficients for SAT were practically unaffected by regression through origin. Similarly, the coefficient estimates from omitting a dummy variable (which used intercept regression) were the difference between the omitted dummy variable's coefficient and the other coefficients in the no intercept results in Table F (within rounding). The "direction" of this effect was inferred by linearly ordering the unstandardized coefficients for the Alt_Dummy variables. As these means increased (al4 increased), Exiting increased, which is consistent with Equation F in Table F. Finally, the

[4] Blalock (1979) points out that multiple categorical variables cannot be jointly estimated using regression through the origin.

aggregated coefficient for the Alt4 dummy variables, "confirmed" the Equation F results that Alternatives, estimated as a categorical variable, was positively associated with Exiting as hypothesized.

However, the proposed approach is tedious to use. The adjusted variance-covariance matrix for Latent Variable Regression must be manually calculated. Similarly, the standard errors for Latent Variable Regression must be manually calculated (measurement model covariances could be substituted for the calculated covariances), and aggregation of the dummy variable coefficient results must be manually performed. (EXCEL templates are available from the authors for these calculations.)

The sample size of each dummy variable was a fraction of the total sample. Although most of the dummy variables were significant in the examples, the typically small sample sizes of survey data tests (e.g., about 200 cases) can produce one or more nonsignificant associations in the dummy variables because of their small subsamples. Thus, aggregation of the dummy variables is desirable to judge the overall significance of the categorical variable from which they were created.[5]

The suggested aggregation approach for the dummy variables was uninvestigated for any bias and inefficiency. Thus, the significance threshold for the aggregated coefficient of a categorical variable probably should be higher than the customary |t-value| = 2.0, where "| |" indicates absolute value. For example, since the significances of the Table E and Table F aggregated coefficients were materially larger than t = 2.0, they were judged to be significant.

[5] Parenthetically, the significance of the Sa2-Exiting association in Table A, for example, was very different from that of the aggregation of its dummy variables shown in Table E because the associations themselves were estimated differently.

Obviously, Latent Variable Regression is limited to a estimating a single dependent or endogenous variable, and it provides Least Squares coefficient estimates, rather than the preferred Maximum Likelihood estimates. However, Castella (1983) has proposed adding a "leverage data point" that "forces" a regression intercept of zero (i.e., the intercept or constant estimated in no-origin regression is zero). This leverage data point also may permit Equation 2, for example, to be estimated using covariant structure analysis and all its dummy variables with maximum likelihood.

Unfortunately, the results for the dummy variables are sensitive to how the dummy variables are coded. Changing the assignment of 1 for category exclusion (e.g., Sat_Dummy5 = 1 if Sa2 = 5, and Sat_Dummy5 = 0 otherwise) to -1, for example, for category inclusion, reverses the signs on all the dummy variables in the tables. However, this sensitivity to coding provides further "meaning" for dummy variables. Specifically, the absolute value of the unstandardized coefficient of Sat_Dummy5, for example, in Table D is the change from zero, the (arbitrary) category exclusion value for the dummy variable coding, for the mean of EXI that is "caused" (associated) with Sat_Dummy5 (.15). In this case, the mean of EXI changed in absolute value (from zero) for the "very satisfied" category by the amount .15. Similarly, the absolute value of the change in EXI (from zero) for the "very dissatisfied" (Sat_Dummy1) was 1.51.

The "direction" of a (truly) categorical association across its categories can be nearly impossible to hypothesize in the customary manner. For example, it is not obvious how one would argue, a priori, the "directionality" of any association between the eight VALS Psychographic categories (SRI International 1989) and Exiting using customary hypotheses such as

H2: VALS increases Exiting

or

H2': VALS decreases Exiting.

However, an hypothesis involving VALS and Exiting might be stated without "directionality" as

H2": VALS is associated with (affects) Exiting.

Any "directionality" could be inferred later from linearly ordering the resulting means (even if they were difficult to explain). Such an approach of disconfirming an association stated without directionality, then observing or "discovering" directionality is within the "logic of discovery" (e.g., Hunt 1983), so long as this "discovery" of directionality is presented as potentially an artifact of the study that must be hypothesized, theoretically supported, then disconfirmed in an additional study.

SUMMARY

The paper suggested an approach to estimating an exogenous (truly) categorical variable (e.g., Gender) in theory (hypothesis) tests of latent variable models with survey data. The approach involved using dummy variables for the categories, and Latent Variable Regression (Ping 1996). The dummy variable estimation results were aggregated to gauge the disconfirmation of a categorical variable hypothesis. The paper also suggested that associations between exogenous categorical variables and endogenous latent variables might be hypothesized without the customary "directionality" statement (that can be difficult to predict in categorical variables).

REFERENCES

Anderson, James C. and David W. Gerbing (1988), "Structural Equation Modeling in Practice: A Review and Recommended Two-Step Approach," *Psychological Bulletin*, 103 (May), 411-23.

Blalock, Hubert M., Jr. (1979), *Social Statistics*, New York: McGraw Hill.

Castella, G. (1983), "Leverage and Regression Through the Origin, The American Statistician, 37 (2) (May), 147-152.

Gordon, H. A. (1981), "Errors in Computer Packages: Least Squares Regression through the Origin," *The Statistician*, 30, 23-29.

Hahn, G. (1971), "Fitting regression Models with no Intercept Term," *Journal of Quality Technology*, 9, 56-61.

Hunt, Shelby D. (1983), *Marketing Theory: The Philosophy of Marketing Science*, Homewood, IL: Richard D. Irwin.

Jaccard, James, Robert Turrisi and Choi K. Wan (1990), *Interaction Effects in Multiple Regression*, Newbury Park, Ca: Sage Publications.

Jöreskog, Karl G. and Dag Sörbom (1996), *Lisrel 8 User's Reference Guide*, Chicago: Scientific Software International, Inc.

Ping, R. (1996), "Latent Variable Regression: A Technique for Estimating Interaction and Quadratic Coefficients," *Multivariate Behavioral Research*, 31 (1), 95-120.

______ (2001), "A Suggested Standard Error for Interaction Coefficients in Latent Variable Regression," *2001 Academy of Marketing Science Conference Proceedings*, Miami: Academy of Marketing Science.

SRI International (1989), *The VALS 2 Segmentation System*, Menlo Park, CA: SRI International.

Table A--Abbreviated Equation 1[a] Estimation Results[b]

Equation 1 Estimation Results:

Endog	EXI		
Exog	Unstd		
Variable	Str Coef	SE[c]	t-value
ALT	0.63	0.11	-5.93
SAT	-0.58	0.10	-5.76

Equation 1 with Sa2[d] Estimation Results:

Exog	Unstd		
Variable	Str Coef	SE[c]	t-value
ALT	0.69	0.07	9.65
SA2	-0.42	0.05	-7.79

[a] $EXI = b_1SAT + b_2ALT + \zeta$.

[b] The structural models were judged to fit the data. Estimates involved LISREL and maximum likelihood.

[c] SE is Standard Error.

[d] $EXI = b_1'Sa2 + b_2'ALT + \zeta''$.

Table B--Abbreviated Equation 2 Results with Sat_Dummy1 Omitted[a b]

Exog Variable	Unstd Str Coef	SE[c]	t-value
ALT	0.68	0.07	9.92
Sat_Dummy3	-0.98	0.26	-3.79
Sat_Dummy4	-1.30	0.26	-5.05
Sat_Dummy2	-0.02	0.26	-0.06
Sat_Dummy5	-1.36	0.28	-4.87

[a] EXI = ~~b_{11}Sat_Dummy1~~ + b_{12}'Sat_Dummy2 + b_{13}'Sat_Dummy3
+ b_{14}'Sat_Dummy4 + b_{15}'Sat_Dummy5 + b_2"ALT + ζ'"

[b] The structural model was judged to fit the data. Estimates involved LISREL and maximum likelihood.

[c] SE is Standard Error.

Table C--Abbreviated Equation 2 Results with Sat_Dummy5 Omitted[a b]

Exog Variable	Unstd Str Coef	SE[c]	t-value
ALT	0.68	0.07	9.92
Sat_Dummy3	0.38	0.15	2.48
Sat_Dummy1	1.36	0.28	4.87
Sat_Dummy4	0.06	0.12	0.51
Sat_Dummy2	1.34	0.18	7.53

[a] EXI = b_{11}'Sat_Dummy1 + b_{12}"Sat_Dummy2 + b_{13}"Sat_Dummy3
+ b_{14}"Sat_Dummy4 + ~~b_{15}Sat_Dummy5~~ + b_2"'ALT + ζ""

[b] The structural model was judged to fit the data. Estimates involved LISREL and maximum likelihood.

[c] SE is Standard Error.

Table D-- Abbreviated Equation 2 Results with all the Sa2 Dummy Variables[a][b]

Exog Variable	Unstd Str Coef	SE[c]	t-value
ALT	0.68	0.07	9.92
Sat_Dummy1	1.51	0.36	4.20
Sat_Dummy5	0.15	0.18	0.85
Sat_Dummy3	0.54	0.24	2.21
Sat_Dummy2	1.50	0.28	5.37
Sat_Dummy4	0.22	0.20	1.09

[a] $EXI = b_{11}''\text{Sat\_Dummy1} + b_{12}'''\text{Sat\_Dummy2} + b_{13}'''\text{Sat\_Dummy3}$
$+ b_{14}'''\text{Sat\_Dummy4} + b_{15}'\text{Sat\_Dummy5} + b_{2}''''\text{ALT} + \zeta'''''$

[b] Estimation involved Latent Variable Regression with least squares.

[c] The standard error (SE) is from Ping (2001).

Table E--Aggregation Results[a] [b]

Case Weighted Average of the Unstandardized Structural Coefficients	Standard Error of the Case Weighted Average of the Unstandardized Structural Coefficients	T-value of Case Weighted Average of the Unstandardized Structural Coefficients
0.51	0.21	2.47

[a] The Case Weighted Average is $\Sigma w_i b_i$, where Σ is summation, $i = 1$ to 5, w_i is the weighted average (number of cases in category i divided by the total number of cases) of the unstandardized coefficient, b_i , of Sat_Dummyi, and i = 1 to 5.

[b] The aggregated Standard Error is the Square Root of the variance of the weighted sum of the individual standard errors (e.g., $\text{sqrt}(\text{Var}(w_1SE_1 + w_2SE_2 + w_3SE_3 + w_4SE_4 + w_5SE_5) = \text{sqrt}(\Sigma w_i^2 SE_i^2 + 2(\Sigma \text{Cov}(SE_i,SE_j)))$, where "sqrt" is the square root, Var is variance, w_i is the weighted average (number of cases in category i divided by the total number of cases) of the unstandardized coefficient of Sat_Dummyi, SE is standard error, Σ is summation, $i = 1$ to 5, $j = 2$ to 5, and $i > j$.

Table F--Abbreviated Estimation Results[a] for ALT as a Categorical Variable

a) Abbreviated Equation 1 with Al4[b] Estimation Results: [c]

Exog Variable	Unstd Str Coef	SE[d]	t-value
SAT	-0.57	0.07	-8.59
AL4	0.52	0.05	9.69

b) Abbreviated Equation 2 Results with Al4[b] and Alt_Dummy1 Omitted [e]

Exog Variable	Unstd Str Coef	SE[d]	t-value
SAT	-0.56	0.07	-8.56
Alt_Dummy3	0.67	0.18	3.85
Alt_Dummy5	1.80	0.26	7.02
Alt_Dummy4	1.56	0.20	7.72
Alt_Dummy2	0.27	0.17	1.61

c) Abbreviated Equation 2 Results with Al4[b] and Alt_Dummy5 Omitted [f]

Exog Variable	Unstd Str Coef	SE[d]	t-value
SAT	-0.56	0.07	-8.56
Alt_Dummy3	-1.12	0.19	-5.83
Alt_Dummy1	-1.80	0.26	-7.02
Alt_Dummy4	-0.23	0.19	-1.21
Alt_Dummy2	-1.53	0.21	-7.42

d) Abbreviated Equation 2 Results with All Al4[b] Dummy Variables[g]

Exog Variable	Unstd Str Coef	SE[d]	t-value
SAT	-0.56	0.07	-8.56
Alt_Dummy5	5.70	0.23	24.45
Alt_Dummy4	5.47	0.23	23.89
Alt_Dummy1	3.90	0.34	11.54
Alt_Dummy3	4.58	0.26	17.90
Alt_Dummy2	4.17	0.29	14.62

Table F (con't.)--Abbreviated Estimation Results[a] for ALT as a Categorical Variable

e) Aggregation Results [h] [i]

Case Weighted Average of the Unstandardized Structural Coefficients	Standard Error of the Case Weighted Average of the Unstandardized Structural Coefficients	T-value of Case Weighted Average of the Unstandardized Structural Coefficients
4.48	0.27	16.86

[a] The structural models were judged to fit the data. Exhibits a) through c) involved LISREL and maximum likelihood estimates, and Exhibit d) involved Latent Variable Regression and least squares estimates.

[b] Al4 was the heaviest loading indicator of ALT.

[c] EXI = b_1''SAT + b_2''Al4 + ζ''''' .

[d] SE is Standard Error.

[e] EXI = b_1'''SAT + ~~b_{21}Alt_Dummy1~~ + b_{22}Alt_Dummy2 + b_{23}Alt_Dummy3
+ b_{24}Alt_Dummy4 + b_{25}Alt_Dummy5 + ζ'''''' .

[f] EXI = b_1''''SAT + b_{21}Alt_Dummy1 + b_{22}'Alt_Dummy2 + b_{23}'Alt_Dummy3
+ b_{24}'Alt_Dummy4 + ~~b_{25}Alt_Dummy5~~ + ζ''''''' .

[g] EXI = b_1'''''SAT + b_{21}'Alt_Dummy1 + b_{22}''Alt_Dummy2 + b_{23}''Alt_Dummy3
+ b_{24}''Alt_Dummy4 + b_{25}''Alt_Dummy5 + ζ''''''''

[h] The Case Weighted Average is $\Sigma w_i b_i$, where Σ is summation, i = 1 to 5, w_i is the weighted average (number of cases in category i divided by the total number of cases) of the unstandardized coefficient, b_i , of Alt_Dummyi, and i = 1 to 5.

[i] The aggregated Standard Error is the Square Root of the variance of the weighted sum of the individual standard errors (e.g., sqrt(Var($w_1SE_1 + w_2SE_2 + w_3SE_3 + w_4SE_4 + w_5SE_5$) = sqrt($\Sigma w_i^2 SE_i^2$ + 2(ΣCov(SE_i,SE_j)))), where "sqrt" is the square root, Var is variance, w_i is the weighted average (number of cases in category i divided by the total number of cases) of the unstandardized coefficient of Alt_Dummyi, SE is standard error, Σ is summation, i = 1 to 5, j = 2 to 5, and i > j.