OFFICE OF SCALE RESEARCH

Technical Report
#0101

Beyond Validity

by

Dr. Gordon C. Bruner II

BEYOND VALIDITY

ABSTRACT

It is not unusual for new scales to be created with little concern for their relationship with previous measures of the same construct (standardization) and little if any reasoning provided for their development (justification). This practice is challenged and the results of three empirical studies illustrate the potential negative consequences this can have on a stream of research.

This paper is more advanced and detailed version of thoughts presented in TR#9901.

Beyond Validity

Usage of multi-item scales has become quite popular over the last couple of decades in conducting market research. The articles and books on the topic attest to the broad acceptance of the approach in both industry as well as academia. What also is clear, however, is that for most constructs there is little agreement about which preexisting scale to use or whether it is just as good to create a new scale for the purpose. A recent study of A_{ad}, for example, concluded that of 75 instances of its measurement, almost half of the scales had been used as a set of items just once (Bruner 1998). In other words, it has not been unusual for new or substantially modified measures to be used despite the existence of other scales. The result is that while authors may be calling what they are measuring the same thing (e.g., A_{ad}), the lack of information about measure equivalency leaves open the possibility that variance in findings across studies is occurring because different scales are being used.

This paper proposes that scholarly marketing research has matured to the point where evidence attesting to a scale's reliability and validity is important but insufficient reasoning for use of a new scale in a study. It is argued that *standardization* and *justification* should be incorporated into the scale selection and usage paradigm. To illustrate the potential negative effects of a lack of standardization, the results of three studies will be presented. Suggestions for improving scale development and usage are also provided.

Background

Standardization

With regard to psychometric scales, standardization can be defined as the extent to which *the same rules* are used to assign numbers to a construct. The use of agreed upon measures is common in most fields that attempt to be scientific. In fact, a fundamental principle of science is that any observation made by one researcher should be independently verifiable by other researchers and this "principle is violated if scientists can disagree about the measure" (Nunnally and Bernstein 1994, p. 6). If a new measure is used and the relationship between it and another established measure is known then that may be acceptable as well, e.g., transformation rules for congeneric tests (Traub 1994, p. 58). But, the point is that standardization is lacking when a variety of different scales are used by researchers to measure the same construct and the equivalency of the scales is unknown.

There are several reasons for standardizing usage. First, some degree of standardization is necessary in order to validate a scale because validation is best viewed as a process; it is not likely to occur in one study but requires the methodical testing of a measure in multiple studies to produce a knowledge base of its psychometric properties (e.g., Cronbach 1971; Peter 1981). Similarly, no single use of a scale is likely to examine all of the facets of generalization that should be addressed by any measure for which wide application is desired (Rentz 1987; Finn and Kayandé 1997). Third, a proper "final" step in scale construction is development of norms (Churchill 1979). This is helpful because scale scores are best interpreted in light of normative data regarding a measure's use with different groups and in a variety of situations (Furse and Stewart 1982). Thus, multiple studies across conditions and populations should be conducted over time utilizing *the same scale* in order to achieve the important psychometric goals of measure validation, generalizability, and norm development.

In contrast, the continuity and value of the research process in a discipline is undermined when scales are developed, used once, and then rarely if ever used again (Chun, Cobb, and French 1975, p. x). This was identified as a problem decades ago in the parent disciplines of marketing. In sociology journals for the period of 1954 to 1965, 72% of scales were found to have been used just once (Bonjean, Hill, and McLemore 1967). Similarly, for the period of 1960 to 1969 in psychology, 63% of scales were used just one time (Chun, Barnowe, Cobb, and French 1974). Although there has been no known broad examination of marketing scale usage, in a study of consumer scales it was estimated that 79% of those published during in the 1980s were used just once (King and Bruner 1993).

Beyond these general reasons for the importance of standardization another concern is that measurement diversity has the potential to affect research results. Alternative measures of the same construct can vary in the amount of a construct's variance that is explained. It is quite possible that a construct has multiple facets and one apparently "good" scale taps more heavily into one facet of interest than another apparently "good" scale with the potential to lead to different findings. For example, in several studies where multiple measures of A_{ad} have been used, the conclusions drawn about the significant relationships of one measure were different from those based on the other measure(s) (e.g., Burton and Lichtenstein 1988; Miniard, Bhatla, and Rose 1990; Olney, Holbrook, and Batra 1991; Stafford 1998). This diversity of operationalization is a possible reason why Brown and Stayman (1992) discovered significant variation across studies in their meta-analysis ad attitudes.

Justification

Justification is defined here as providing *adequate reasoning* for the use of a particular measure. One type of justification is evidence of psychometric quality. This is most called for when a new scale is used. We have come to expect that scale creators at least provide evidence of unidimensionality and internal consistency if not for convergent and discriminant validity as well (e.g., Churchill 1979; Gerbing and Anderson 1988). Providing this sort of evidence is also important for measures that are modified in some way. In the field of psychology, a primary standard of measure use (as opposed to a conditional standard) is that it is incumbent upon subsequent users of a measure who modify it in some way to revalidate it or at least explain why additional validation is unnecessary (AERA, APA, and NCME 1985, p. 41).

Even when an established scale is used some limited justification is helpful to provide. It may be as simple as citing the relevant sources where evidence of the scale's validity can be found. When such information is not provided readers must guess whether the scale is new, adapted, or borrowed as well as the degree to which it is a valid measure and how it relates to previous measures. This uncertainty has the potential to decrease confidence in the findings associated with the scale.

Another type of justification is called for when a new scale is developed even though others are available. Researchers should use previously developed measures unless they can explain why it is not possible or appropriate (Churchill 1979; Varadarajan 1996). One reason researchers may not have done this in the past is because they were simply unaware that appropriate scales were already available. This has become a less defendable excuse in the last decade, however, due to the ability to easily conduct computerized searches of Internet or other digitized databases. Further, books totally devoted to the topic have been published within marketing (Bearden and Netemeyer 1998; Bruner, James, and Hensel 2001) as well as related fields (e.g., McDowell and Newell 1996; Robinson, Shaver, and Wrightsman 1999; Mental Measurements Yearbook 2001).

Some reasons for not using pre-existing scales are defendable. For example, the available scales may be considered to be too complex or lengthy for a usage situation. Yet other possibilities are that available scales are viewed as having weaknesses in content or construct validity. Beyond those reasons, however, is the view that there is no particular problem producing an additional measure of a construct even when alternatives are known and considered adequate. It can be argued that if a new scale's items have been drawn from the same semantic domain as other measures they can be assumed to have a similar amount of common core. In other words, "parallel" scales can be developed and lead to essentially the same conclusions when used in empirical research.[1]

Although theoretically appealing, the strength of this view depends upon the degree to which the items composing two or more scales were *randomly sampled* from the same well defined domain. In reality, it is highly unlikely that such a process has been used in our field. Instead, items have been *constructed* which, at best, produce "alternate form" scales with unknown statistical equivalence. This introduces various forms of errors, notably systematic differences, that are not covered by the domain sampling model (Nunnally and Bernstein 1994, p. 250, 251).

Empirical Illustrations

As stated above, standardization and justification may be considered unnecessary if one assumes that the domain sampling model fits the situation. This assumption leads some researchers to think that if a scale appears on the surface to tap into the same semantic domain as a preexisting scale then it is unnecessary to provide confirming evidence of that judgment and use of either scale in the study would produce substantially the same results. The fallibility of this assumption is demonstrated here as simply as possible by showing the consequences of using different scales to measure the same construct.[2]

Study 1

The purpose of this study was to examine several antecedents of the advertising hierarchy of effects as it relates to websites. Among the antecedents examined was time spent on the Internet. It was expected that familiarity and experience using the Internet affects how stimuli on the web are processed.

Two different A_{ad} scales with mutually exclusive sets of items were employed. One set of items (A_{ad1}) was derived from Mitchell and Olson (1981) and has been used by many other (e.g., Miller and Marks 1992; Stafford and Day 1995). The second scale (A_{ad2}) was taken from Muehling, Stoltman, and Misra (1990). This set of items is a popular nucleus that has been augmented with other items in various studies (e.g., Muehling and Laczniak 1992; Zinhan and Zinkhan 1985). (Items for all scales as well as their respective reliabilities are shown in the Appendix.)

The sample could be generally described as non-college students. Ninety-five adults took part in the study with about sixty percent of the sample being female and the largest age group being between 40 and 59 (73%).

A simple ANOVA can be used to illustrate the problem. As shown in Table 2, level of internet usage clearly was related to attitude-toward-an-ad shown viewed at a website (p = .01) when A_{ad1} was used. However, when A_{ad2} was used the conclusion would be that there was no significant effect.

[Place Table 1 about here]

Study 2

The second study was part of an experiment noting the effects of two different websites on various perceptual and attitudinal constructs. Among the constructs measured was attitude-toward-the-website (A_{ws}). This is a relatively new construct that is likely to become as popular in the future as A_{ad} has been in the past.

Only two scales of A_{ws} were noted in the literature at the time: a six-item scale by Chen and Wells (1999) and a three-item version by Bruner and Kumar (2000). Although their items are very different both have been presented in the literature as Likert-type global measures of the same construct.

This experiment's sample was college students almost evenly split on gender (51% female). The majority were in their early 20s (81% were 20 or 21). About half of the subject's were exposed to one website while the other half were exposed to the second site. As in Study 1, a simple ANOVA illustrates the problem (Table 2). The use of A_{ws2} would lead a researcher to believe that there was a significant difference in subjects' attitudes toward the two different sites. However, use of A_{ws1} would not lead most researchers to that same conclusion because the level of significance was much weaker (beyond the typical .05 level).

[Place Table 2 about here]

Study 3

A somewhat more elaborate illustration is provided in this study by showing the compounding effects that occur when an analysis employs more than one scale. The experiment was designed to compare the effects of different visual backgrounds on a variety of viewer responses to an advertisement at a website. Specifically, it was anticipated that an advertisement in the context of a complex background would be associated with poorer A_{ad} and A_{b} than simpler backgrounds.

The sample was composed of college students randomly assigned to three treatments. A little over half of the sample was male (59%) and 54% were between the ages of 20 and 24.

Included in the questionnaire were two mutually exclusive sets of items for measurement of A_{ad} and A_{b}. The two versions of A_{ad} were the same as described for Study 1. Regarding A_{b}, version one has been used by Alpert and Kamins (1995) as well as Miniard, Bhatla, and Rose (1990). Among others, the second version of A_{b} has been used by Loken and Ward (1990) as well as Ward, Bitner, and Barnes (1992).

The effect of webpage background (independent variable) on A_{ad} and A_{b} (dependent variables) was tested by means of a multivariate analysis of variance (MANOVA) as well as complementary ANOVAs. The results are presented in Table 3. A total of four MANOVAs were run, each using a different pair of A_{ad} and A_{b} scales to determine the extent to which the results and conclusions might vary.

[Place Table 3 about here]

The results of the ANOVAs indicate that A_{ad1} and A_{b1} were significantly influenced by the independent variable (complexity of webpage background) whereas A_{ad2} and A_{b2} were not. This, in turn, led to the varying results of the MANOVAs. Depending upon the set of scales used, conclusions could have ranged from full support for the hypothesized relationships, to partial support, to no support.

Discussion

Having one theoretical conceptualization for our primary constructs (e.g., A_{ad} and A_b) as well as one set of accepted scales would go a long way towards resolving problems of inconsistent findings such as illustrated above. However, that is an unrealistic goal for the near future for most of our constructs. Alternative views are likely to coexist for some time as will a variety of measures related to each conceptualization.

In the meantime, the procedure outlined in Table 4 is offered as a guide for implementing standardization and justification in the selection and development of scales. The emphasis is on effort to identify and use previously developed scales. Development of new scales should be conducted only if necessary and carries with it certain validation and reporting responsibilities.

[Place Table 4 about here]

Faith in the domain sampling assumption should be tempered by the realization that the *judgment* of researchers of what the domain is and what constitutes an adequate set of items from that domain will differ. The burden of evidence is on developers of a new scale, not reviewers or readers; authors must explain why existing scales are inadequate and can not be used.

Lack of unidimensionality would indicate that two sets of items are tapping into different domains. However, evidence of unidimensionality appears to be insufficient as well. In each of the studies described above, the factor analyses indicated that the items composing the pairs of scales were unidimensional yet, use of the separate item subsets led to different conclusions! This highlights the fact that casual judgments and simple testing (e.g., face validity, internal consistency, unidimensionality) are not always adequate to determine of whether scales are similar enough so as to result in the same conclusions. More rigorous testing is required.

It is acknowledged that average scale users are unlikely to go to the effort of providing ample evidence of reliability, validity, and equivalency (with previous measures) for every scale they use. They would rather get on with testing relationships. Given this, our field would benefit from more articles that critically compare competing scales that have been developed for popularly measured constructs. Comparisons should be rigorous enough to gauge the equivalency of the scales ranging from being *parallel tests* on the high end to *congeneric tests* on the low end (e.g., Traub 1994, Ch. 5.) Of course, it is also possible that such testing could indicate that the competing scales are actually measuring different, though related constructs. If possible, these articles should do such testing as necessary to allow conclusions to be drawn about the overall superiority of a scale or at least its relative superiority for certain contexts. Other researchers could then refer to these articles as they make choices among scales. Such articles could also note areas for future research such as when all available scales have serious limitations.

Short of having the results of such empirical comparisons to guide us, there should be a presumption of difference between scales. This is in contrast to what seems to be occurring currently where as long as a researcher uses the same name for a measure as used by previous researchers and provides some limited evidence of its reliability then the scale and its associated findings are accepted with little question. Due to this it is recommended that reviews, meta-analyses, and syntheses of findings across studies be conducted more cautiously. If the scales used in reviewed studies are the same or very similar then comparison of results may be safe. In contrast, comparison of findings across studies with different scales should be sensitive to the strong possibility that disparate conclusions may have been reached by the various researchers due to the method variance of the type described in this paper.

FOOTNOTES

1. Evidence of this view is provided in a special appendix for the editor and reviewers only.
2. It is unlikely for most researchers to use two or more multi-item scales to measure the same construct in a study and thus they would not be aware of the potential variance in conclusions that can result. Instead, what is presented here is more akin to what would happen if the same relationships were being examined over time in studies conducted by independent researchers who utilized different measures with little concern about standardization and justification.

Table 1

The Effect of Internet Usage on A_{ad}

			MEANS			
Scale	**F-Ratio**	**Sig. Level**	**None**	**Minutes**	**1 Hour**	**Hours+**
A_{ad1}	4.029	.010	4.060	4.536	5.355	4.962
A_{ad2}	1.618	.191	3.972	4.222	4.865	4.641

Table 2

Website Differences in Aws

			MEANS	
Scale	**F-Ratio**	**Sig. level**	**Site 1**	**Site 2**
Aws1	11.512	.001	4.180	3.764
Aws2	3.127	.081	3.692	3.504

Table 3

The Effect of Webpage Background on Ad Effectiveness

			MEANS		
Variable	F-Ratio	Sig. level	Webpage1	Webpage2	Webpage3
Multivariate test					
(Wilk's lambda)	2.496	.045			
Univariate tests					
A_{ad1}	3.474	.035	5.525*	5.158	4.650
A_{b1}	3.987	.022	4.911*	4.700	4.200
Multivariate test					
(Wilk's lambda)	1.939	.106			
Univariate tests					
A_{ad1}	3.474	.035	5.525*	5.158	4.650
A_{b2}	2.219	.115	4.711	4.611	4.200
Multivariate test					
(Wilk's lambda)	2.099	.083			
Univariate tests					
A_{ad2}	1.662	.196	5.411	5.033	4.744
A_{b1}	3.987	.022	4.911*	4.700	4.200
Multivariate test					
(Wilk's lambda)	1.406	.234			
Univariate tests					
A_{ad2}	1.662	.196	5.411	5.033	4.744
A_{b2}	2.219	.115	4.711	4.611	4.200

* In post-hoc comparisons (Tukey HSD) the means associated with Webpage1 were found to be statistically different ($p \leq .05$) from those associated with Webpage3. None of the means associated with Webpage1 were statistically different from those associated with Webpage2.

Table 4

SCALE SELECTION & USAGE PROCESS

1. Determine the latent construct to be measured.

2. Determine if a multi-item scale is an appropriate *type* of measure for empirically operationalizing a construct. A ratio level measure may be possible and preferable while in other cases multi-item scales may be not be possible due to the survey length/time restrictions.

3. If a multi-item scale is appropriate, search to determine if an acceptable scale exists. Conduct computerized searches of Internet or other digitized databases as well as looking in the scales handbooks.

4. If more than one alternative scale is found, compare them using several criteria:

    a. Face validity - does the scale appear to capture the meaning one intends? It is quite possible that a scale is called one thing by one researcher but is referred to as something else by another researcher.

    b. Psychometric quality - what support is available attesting to the scale's unidimensionality, reliability, validity, as well as norms for use with different samples?

    c. Typicality/acceptance - researchers should have an understanding of the alternative views of the construct that may exist and the alternative scales linked to those views. Choice of a scale should be influenced by knowledge of what scales were used by previous researchers on whose work the current researchers hope to build.

d. Once the scale has been selected and used in a study, a paper will be written. A minimal amount of information about the scale should be provided:

   i. Source – the origin of the scale should be specified. It may also be helpful to cite a recent use or two of the scale in the field if the scale was developed many years earlier or originated in another field.

   ii. Reliability – an estimate of the scale's reliability should be reported. While measuring stability (test-retest) may be popular in other fields (Litwin 1995, p. 8), internal consistency is overwhelmingly the most prevalent type reported in marketing (Peterson 1994, p. 382).

   iii. Indications of any substantial changes that were made. If items were added or deleted, particularly if the result of factor analysis, then that should be described. Rephrasing of items and use a different response format (points, anchors) would be helpful to report as well.

5. If no scale is found or those found are unacceptable then develop one using widely accepted procedures (e.g., Churchill 1979; Gerbing and Anderson 1988).

   a. When the research is reported, some rationale for the scale's construction should be provided, e.g., no known scale was available or those available were inadequate in some specific way.

   b. It's incumbent upon the scale developer to provide more information about the scale's content and psychometric quality in this case than in the case of #4d (above).

i. The scale items, response format (points, anchors), and evidence of unidimensionality, reliability, and validity should be provided.

ii. If the new scale is being offered as an alternative to an established scale the equivalency of the two should be tested.

iii. If such information can not be included with the published article due to length restrictions, an appendix should be made available directly from the authors or from the publishing journal’s website.

APPENDIX

Characteristics of Scales Examined

Attitude-Toward-the-Ad (A_{ad1}), $\alpha = .91$ & $.87$*, (e.g., Mitchell and Olson 1981)

1. good/bad
2. like/dislike
3. irritating/not irritating
4. interesting/uninteresting

Attitude-Toward-the-Ad (A_{ad2}), $\alpha = .91$ & $.83$*, (Muehling, Stoltman, and Misra 1990)

1. unappealing/appealing
2. unattractive/attractive
3. unpleasant/pleasant

Attitude-Toward-the-Brand (A_{b1}), $\alpha = .89$, (e.g., Miniard, Bhatla, and Rose 1990)

1. dislike/like
2. unfavorable/favorable
3. negative/positive

Attitude-Toward-the-Brand (A_{b2}), $\alpha = .91$, (e.g., Loken and Ward 1990)

1. bad/good
2. poor quality/high quality
3. unsatisfactory/satisfactory

Attitude-Toward-the-Website (A_{ws1}), $\alpha = .75$, (Chen and Wells 1999)

1. This website makes it easy for me to build a relationship with this company.

2. I would like to visit this website again in the future.

3. I'm satisfied with the service provided by this website.

4. I feel comfortable in surfing this website.

5. I feel surfing this website is a good way for me to spend my time.

6. Compared with other websites, I would rate this website as *one of the worst/one of the best*.

Attitude-Toward-the-Website (A_{WS2}), α = .91, (Bruner and Kumar 2000)

1. I liked the website.

2. I think it is a good website.

3. I think it is a nice website.

* The alphas calculated for studies 1 and 3, respectively.

Reviewers' Appendix

These quotes are not intended for the final paper (if accepted). They are provided to reviewers of the current paper to substantiate my description of the confidence that some researchers and reviewers place in the domain sampling model. The statements are from four different reviewers regarding other papers written several years ago, not the one you are reviewing.

" . . . domain sampling theory suggests that different items may be sampled from the specific (core) domain for a given construct without having a substantial effect on tests of proposed relationships involving the measure . . . . If items generally seem to fit the conceptualization and inter-item correlations are high, I doubt that we have seen a large number of misleading conclusions in the literature."

"Given that A_{ad} *is often conceptualized as a person's global evaluation of the ad, just as* A_{o} *or* A_{brand} *is conceptualized as a person's global evaluation of the brand, both construct's are operationalized quite commonly with semantic differential attitude scales. The validity of such scales was established long ago in Osgood, Suci, and Tannenbaum's (1957) classic work. Thus, so long as one conceptualizes* A_{ad} *as an attitude, one can simply adopt this standard attitude measurement, just as is done for* A_{o}*. There is no need for the Churchill paradigm in this situation. It is an irrelevant benchmark."*

"Given the assumption that there exists a domain of scale items that are representative of a construct, measures of A_{ad} *could be different from one another, yet still be valid representations of the same construct, depending on which items the researcher sampled from the domain."*

"If the measures are not tapping into the same semantic domain, what domain are they tapping? Why would authors use the scale items if they did not believe them to be measures of A_{ad}*?"*

References

Alpert, Frank H. and Michael A. Kamins (1995), “An Empirical Investigation of Consumer Memory, Attitude, and Perceptions Toward Pioneer and Follower Brands,” ***Journal of Marketing***, 59 (October), 34-45.

American Educational Research Association (AERA), American Psychological Association (APA), & National Council on Measurement in Education (NCME) (1985), ***Standards for Educational and Psychological Testing***. Washington, D.C.: American Psychological Association.

Bearden, William O. and Richard G. Netemeyer (1998), ***Handbook of Marketing Scales***, second edition. Thousand Oaks, CA: Sage Publications.

Berlyne, D. E. (1960), ***Conflict, Arousal, and Curiosity***. New York: : McGraw-Hill.

Bonjean, Charles M., Hill, Richard J., and McLemore, S. Dale (1967), ***Sociological Measurement: An Inventory of Scales and Indices***. San Francisco: Chandler Publishing Company.

Brown, Steven P. and Douglas M. Stayman (1992), “Antecedents and Consequences of Attitude Toward the Ad: A Meta-Analysis.” ***Journal of Consumer Research***, 19 (June), 34-51.

Bruner II, Gordon C. (1998), “Standardization & Justification: Do Aad Scales Measure Up?” ***Journal of Current Issues & Research in Advertising***, 20 (Spring), 1-18.

----, Karen E. James, and Paul, J. Hensel (2001), ***Marketing Scales Handbook- V. III***. Chicago: American Marketing Association.

---- and Anand Kumar (2000), "Web Commercials and Advertising Hierarchy-of-Effects," ***Journal of Advertising Research***, 40 (January/April), 35-42.

Buros, Oscar Krisen (1975), ***Personality Tests and Reviews II***. Highland Park: The Gryphon Press.

Burton, Scot and Donald R. Lichtenstein (1988), "The Effects Ad Claims and Ad Content on Attitude Toward the Advertisement," ***Journal of Advertising***, 17 (Spring), 3-11.

Chen, Qimei and William D. Wells (1999), “Attitude Toward the Site,” ***Journal of Advertising Research***, 39 (Sept./Oct), 27-37.

Chun, Ki-Taek, J. T. Barnowe, Sidney Cobb, and John R. P. French, Jr. (1974), ***Publication and Uses of Psychological Measures in the 1960s***. Ann Arbor, Michigan: Institute for Social Research, University of Michigan.

----, Sidney Cobb, and John R. P. French, Jr. (1975), ***Measures for Psychological Assessment***. Ann Arbor, Michigan: Institute For Social Research, University of Michigan.

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

Cronbach, Lee J. (1971), "Test Validation," in ***Educational Measurement***, R. L. Thorndike, ed. Washington, D.C.: American Council on Education, 443-507.

Finn, Adam and Ujwal Kayandé (1997), "Reliability Assessment and Optimization of Marketing Measurement," ***Journal of Marketing Research***, 34 (May), 262-275.

Furse, David H. and David W. Stewart (1982), "Standards For Advertising Copytesting: A Psychometric Interpretation," ***Journal of Advertising***, 11 (Winter), 30-38, 76.

Gerbing, David W. and James C. Anderson (1988), "An Updated Paradigm for Scale Development Incorporating Unidimensionality and Its Assessment," ***Journal of Marketing Research*** 25 (May), 186-192.

King, Maryon F. and Gordon C. Bruner II (1993), "Evaluation of Consumer-Related Scales: A Structured Schemata Approach," ***Office of Scale Research Technical Report #9301.***

Litwin, Mark S. (1995), ***How to Measure Survey Reliability and Validity***. Thousand Oaks, CA: Sage Publications, Inc.

Loken, Barbara and James Ward (1990), "Alternative Approaches to Understanding the Determinants of Typicality," ***Journal of Consumer Research***,17 (September), 111-126.

McDowell, Ian and Claire Newell (1996), ***Measuring Health: A Guide to Rating Scales and Questionnaires***. New York: Oxford University Press.

Mental Measurements Yearbook, 14th edition (2001), James C. Impara, ed. Lincoln, NE: Buros Institute, University of Nebraska Press.

Miller, Darryl W. and Lawrence J. Marks (1992), "Mental Imagery and Sound Effects in Radio Commercials," ***Journal of Advertising***, 21 (Winter), 83-93.

Miniard, Paul W., Sunil Bhatla, and Randall L. Rose (1990), "On the Formation and Relationship of Ad and Brand Attitudes: An Experimental and Causal Analysis," ***Journal of Marketing Research***, 27 (August), 290-303.

Mitchell, Andrew A. and Jerry C. Olson (1981), "Are Product Attribute Beliefs the Only Mediator of Advertising Effects on Brand Attitude?" ***Journal of Marketing Research***, 18 (August), 318-332.

Muehling, Darrel D. and Laczniak, Russell N. (1992), "An Examination of Factors Mediating and Moderating Advertising's Effect on Brand Attitude Formation," ***Journal of Current Issues and Research in Advertising***, 14 (Spring), 23-34.

----, Jeffrey J. Stoltman, and Sanjay Mishra (1990), "An Examination of the Cognitive Antecedents of Attitude-Toward-the-Ad," ***Current Issues and Research in Advertising***, 12 (Spring & Fall), 95-117.

Nunnally, Jum C. and Bernstein, Ira H. (1994), ***Psychometric Theory***. New York: McGraw-Hill.

Olney, Thomas J., Morris B. Holbrook, and Rajeev Batra (1991), "Consumer Responses to Advertising: The Effects of Ad Content, Emotions, and Attitude toward the Ad on Viewing Time," ***Journal of Consumer Research***, 17 (March), 440-453.

Peter, J. Paul. (1981), "Construct Validity: A Review of Basic Issues and Marketing Practices," ***Journal of Marketing Research***, 18 (May), 133-145.

Peterson, Robert A. (1994), "A Meta-Analysis of Cronbach's Coefficient Alpha," ***Journal of Consumer Research***, 21 (September), 381-391.

Rentz, Joseph O. (1987), "Generalizability Theory: A Comprehensive Method for Assessing and Improving the Dependability of Marketing Measures," ***Journal of Marketing Research***, 24 (February), 19-28.

Robinson, John P., Phillip R. Shaver, and Lawrence S. Wrightsman (1999), ***Measures of Political Attitudes***. San Diego, CA: Academic Press.

Stafford, Marla Royne (1998), "Advertising Sex-Typed Services: The Effects of Sex, Service Type, and Employee Type on Consumer Attitudes," ***Journal of Advertising***, 27 (2), 65-81.

---- and Ellen Day (1995), "Retail Services Advertising: The Effects of Appeal, Medium, and Service," ***Journal of Advertising***, 24 (Spring), 57-71.

Traub, Ross E. (1994), ***Reliability for the Social Sciences: Theory and Applications***. Thousand Oaks, CA: Sage Publications.

Varadarajan, P. Rajan (1996), "From the Editor: Reflections on Research and Publishing," ***Journal of Marketing***, 60 (October), 3-6.

Ward, James C., Mary Jo Bitner, and John Barnes (1992), "Measuring the Prototypicality and Meaning of Retail Environments," ***Journal of Retailing***, 68 (Summer), 194-220.

Zinkhan, George M., and Christian F. Zinkhan (1985), "Response Profiles and Choice Behavior: An Application to Financial Services," ***Journal of Advertising***, 14 (Fall), 39-44, 51, 66.