

# **OFFICE OF SCALE RESEARCH**

## **Technical Report #9901**

### **The Effects of Nonstandardization in Scale Usage**

by

Dr. Gordon C. Bruner II

# **THE EFFECTS OF NONSTANDARDIZATION IN SCALE USAGE**

## **Abstract**

Two of the most studied constructs in scholarly marketing research are Attitude-toward-the-ad (Aad) and attitude-toward-the-brand (Ab) but there is little agreement about how they should be measured. It has been common for new or modified measures to be created with little concern for the scale's relationship with previous measures of the same construct and little if any justification provided for the new set of items. Yet, this lack of standardization is not viewed as a serious problem by many researchers and reviewers. It is believed that if two scales appear to measure the same thing then we can assume they will produce similar results. This belief is challenged in this paper and the results of an empirical study are used to show that when different measures of the same thing are used then different conclusions are possible.

## **THE EFFECTS OF NONSTANDARDIZATION IN SCALE USAGE**

Attitude-toward-the-ad (Aad) and attitude-toward-the-brand (Ab) have been studied with great frequency over the last two decades. In fact, it appears that they are the most measured constructs in scholarly marketing research, at least when it comes to the use of simple summated scales (Bruner and Hensel 1996, pp. 98-100, 818-825). Unfortunately, it also appears that there is little agreement about how the constructs should be measured. A recent study of Aad, for example, concluded that of 75 instances of Aad scale usage, almost half of them had been used as a set of items just once (Bruner 1998). In other words, it has been common for many authors to create new or modified measures with little concern for the scale's relationship with previous measures of the same construct and little if any justification for the new set of items. While an intensive of analysis of Ab has not been conducted, it seems that it is experiencing a similar fate (see Bruner and Hensel 1996, pp. 98-100).

As obvious as it is that Aad and Ab are characterized by nonstandardization, many researchers do not appear to view it as a problem. This is evidenced by the continued use of new item sets and the acquiescence of reviewers and editors that allows these scales to be acceptable for publication in top journals. The result is that while authors may be calling what they are measuring the same thing (Aad or Ab), the lack of standardization and justification leaves open the possibility that usage of one set of items could be leading to different conclusions compared to what would have occurred if another set of items was used in the same study.

The primary goal of this study is to illustrate the potential effects of using different measures of what appears to be the same thing. Specifically, two different measures of Aad

and two measures of Ab are used in the same study to note what differences they might make in the conclusions.

## **Background**

A recent analysis of Aad scales charged that the variety of measures used over the years have not had as much in common as their names might suggest (Bruner 1998). This was concluded for three main reasons. First, there has been no one accepted theory of attitude structure (Eagly and Chaiken 1993). This has lead, therefore, to some researchers treating an attitude as a multi-dimensional construct while others considered it to be unidimensional. Second, even within a single conceptualization of Aad or Ab there has been a surprising lack of common scale usage (standardization). New variations in Aad and Ab measurement proliferate every year. (Compare Bruner and Hensel 1992 and 1996). Thirdly, as important as Aad or Ab are in our field, justifying our measures of them does not appear to be a priority, even in our top journals. It is not unusual for many apparently acceptable alternative measures to be ignored in favor of new scales being introduced with little, if any, published reasoning or validation.

Why is it that standardization and justification are not high priorities when study methodology is planned? It could be that standardization and justification are considered unnecessary if a new scale's items have been drawn from the same semantic domain as alternative measures (Appendix). Given this view, it is assumed that one scale composed of items from a semantic domain would not lead to substantially different findings compared to another set of items drawn from the same domain. The crux of the matter, therefore, lies in the extent to which different authors who say they are measuring the same construct have

indeed sampled properly from the same semantic domain. Evidence for this adequacy of sampling is rarely provided, again, apparently because of the overriding assumption by reviewers that when authors state that they are measuring a commonly measured construct it is safe to depend upon their judgement in properly measuring that construct. In other words, if a set of items appears on the surface to tap into the same semantic domain (face validity) as other sets of items used previously then it is unnecessary to provide confirming evidence of that judgement.

The purpose of this study is to question this assumption by empirically examining the consequences of using different scales to measure the same construct. Specifically, in the course of conducting an experiment two measures of Aad and two measures of Ab will be employed. This is a conservative test since the measures are not maximally different. All are semantic differentials with the same response format and have been used before.

### **Methodology**

The context in which the comparison of scales occurred was an experiment testing the effects of different visual backgrounds on Aad and Ab when a commercial is run on just a portion of webpage. The backgrounds were manipulated in terms of their complexities as confirmed by a pretest. Based on studies in psychology (Brown and Farha, 1966; Day, 1967) and with print ads (Morrison and Dainoff 1972) it was expected that a commercial run in the context of a complex background would be associated with poorer Aad and Ab than simpler backgrounds.

Subjects for the experiment were students attending a major midwestern U.S. university. Subjects were randomly assigned such that there were thirty for each of the

three treatments. A little over half of the sample was male (59%), 54% were between the ages of 20 and 24, and the majority (83%) were undergraduates.

The three experimental treatments were prepared with the complexity of their backgrounds being the only variable that was intentionally manipulated. Following the general recommendations by Berlyne (1960, p. 38, 39), complexity was manipulated by adding items, color, and movement.

Subjects completed a brief questionnaire on-line after exposure to the stimuli. Among the measures they completed were items intended to measure Aad and Ab. All of these items have a history of usage in various forms as can be noted from the listings provided by Bruner and Hensel (1996, pp. 98-100, 818-825).

Two scales were selected for the measurement of Aad and Ab (a total of four scales) that have been used previously. With regard to the measurement of Aad, one set of items was derived from Mitchell and Olson (1981) and has been used by several others (e.g., Miller and Marks 1992; Stafford and Day 1995). The second Aad scale 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).

Regarding Ab, version one has been used most recently by Alpert and Kamins (1995) but appears to have been used first by Miniard, Bhatla, and Rose (1990). Among others, the second version of Ab has been used by Loken and Ward (1990) as well as Ward, Bitner, and Barnes (1992). The items composing all of these scales and their respective reliabilities as estimated in this study are reported in Table 1. Note that the

two measures of A<sub>ad</sub> as well as the two measures of A<sub>b</sub> do not have any items in common but do share similarly high reliabilities.

[Place Table 1 about here]

### **Findings**

The effect of webpage background (independent variable) on A<sub>ad</sub> and A<sub>b</sub> (dependent variables) was tested by means of a multivariate analysis of variance (MANOVA) as well as complementary ANOVAs. The results are presented in Table 2. Four MANOVAs were run, each using a different pair of A<sub>ad</sub> and A<sub>b</sub> scales to determine the extent to which the results and conclusions might be different.

[Place Table 2 about here]

Note in the results of the ANOVAs how A<sub>ad1</sub> and A<sub>b1</sub> were significantly influenced by the complexity of webpage background whereas A<sub>ad2</sub> and A<sub>b2</sub> were not. This, in turn, led to the different results of the MANOVAs. Specifically, in the case of the first MANOVA where A<sub>ad1</sub> and A<sub>b1</sub> were used, the test was significant. This would lead to the clear conclusion that webpage background was having a substantial effect on viewers' attitudes towards the commercial they saw on the page as well as on the brand being promoted.

Use of the other sets of scales would appear to have led to somewhat different conclusions. If the typical .05 significance level was set, then none of the rest would have shown conclusively that webpage background was associated with changes in brand

and ad attitudes. While some authors might have argued that the second and third MANOVAs provided limited support for their hypotheses, that would not have been possible for the fourth MANOVA where A<sub>ad2</sub> and A<sub>b2</sub> were used. The lack of significance of the Wilk's lambda (.234) as well as the significance levels greater than .10 for the individual ANOVAs would be interpreted as providing no evidence at all that webpage background had a significant relationship with A<sub>ad</sub> and A<sub>b</sub>.

### **Discussion**

The purpose of this very focused study has been to show that when different measures of supposedly the same construct are used then very different results can result. These differing findings then may lead to different conclusions, i.e., depending upon what measures are used different conclusions may legitimately be drawn.

Admittedly, it would have been possible to find several A<sub>ad</sub> scales and several A<sub>b</sub> scales that could have led to similar conclusions. However, it is stressed that the scales used in this experiment were not some odd set of items selected merely to prove a point; the scales have been used in the literature multiple times over several years as acceptable measures of the same constructs.

To the extent that this approach is common in our field, one could question our sincere desire to become scientific. This is because 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). Such activities in marketing research are thwarting attempts to understand phenomena.

Faith in the domain sampling assumption should be tempered by the realization



that the judgement of researchers of what the domain is and what is an adequate sampling of that domain will differ. The responsibility lies with the researcher to provide evidence that a new scale is indeed measuring the same thing as other scales rather than something else (validation). Thus, the position taken here is that **there should be a presumption of difference unless sufficient evidence is provided to the contrary**. This is in contrast to what appears to be occurring now where as long as the researcher refers to a scale as a measure of A<sub>ad</sub> or A<sub>b</sub> and provides some limited evidence of its reliability then reviewers, editors, and readers accept the scale and its associated findings.

The results of this review show that it is not enough to provide evidence of reliability. Even the provision of some validity evidence is insufficient justification for a new scale when previously developed scales were ignored in favor of constructing a new one especially when no clear reasoning for the new scale was provided.

While one theoretical conceptualization of A<sub>ad</sub> and A<sub>b</sub> as well as one set of scales would go a long way towards resolving problems of inconsistent findings it is an unrealistic goal for the near future. Instead, alternative views are likely to coexist for some time as will a variety of measures related to each conceptualization. In the meantime, researchers should have an understanding of the alternative views and then make an informed choice between them based upon the theoretical network to which they expect their study to contribute. This should be plainly stated in the published version of the paper.

A related suggestion is that researchers must be much more sensitive about trying to identify and select from among the scales used previously in work upon which they are building. Surely a previously developed scale can be found that is suitable for the study's purposes and has evidence of reliability and unidimensionality if not validity too. Likewise, it

is recommended that manuscript guidelines place greater emphasis on measurement justification. New or greatly modified measures should be accompanied with both the reason(s) why the available alternatives were not satisfactory as well as some initial evidence of validity.

Finally, the results of this research suggest that reviews, meta-analyses, and syntheses of findings across studies should be conducted carefully. If scales are the same or very similar then comparison may be safe. In contrast, comparison of findings across studies with scales that do not share any items in common is not encouraged and may very well lead to erroneous conclusions.

## 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.
- Berlyne, D. E. (1960), *Conflict, Arousal, and Curiosity*, New York: McGraw-Hill.
- Brown, Larry P. and William Farha (1966), "Some Physical Determinants of Viewing Time of Three Instructions Sets," *Perception and Psychophysics*, 1 (January), 2-4.
- Bruner II, Gordon C. (1998), "Standardization & Justification: Do Aad Scales Measure Up?" *Journal of Current Issues & Research in Advertising*, 20 (Spring), 1-18.
- \_\_\_\_\_ and Paul J. Hensel (1992), *Marketing Scales Handbook - V. I*, Chicago: American Marketing Association.
- \_\_\_\_\_ and \_\_\_\_\_ (1996), *Marketing Scales Handbook - V. II*, Chicago: American Marketing Association.
- Day, Hy (1967), "The Evaluation of Suggestive Complexity, Pleasantness, and Interestingness for a Series of Random Polygons Varying in Complexity," *Perception and Psychophysics*, 2 (July), 281-286.
- Eagly, Alice H. and Shelly Chaiken (1993), *The Psychology of Attitudes*, Fort Worth, TX: Harcourt, Brace, Jovanovich College Publishers.
- Loken, Barbara and James Ward (1990), "Alternative Approaches to Understanding the Determinants of Typicality," *Journal of Consumer Research*, 17 (September), 111-126.
- 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.
- Morrison, Bruce John and Marvin J. Dainoff (1972), "Advertisement Complexity and Looking Time," *Journal of Marketing Research*, 9 (November), 396-400.

Muehling, Darrel D. and Russell N. Laczniak (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 Ira H. Bernstein (1994), *Psychometric Theory*, New York: McGraw-Hill.

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

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.

**Table 1**  
**Characteristics of Scales Examined**

Attitude-Toward-the-Ad (Aad1),  $\alpha = .91$ , (e.g., Mitchell and Olson 1981)

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

Attitude-Toward-the-Ad (Aad2),  $\alpha = .91$ , (Muehling, Stoltman, and Misra 1990)

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

Attitude-Toward-the-Brand (Ab1),  $\alpha = .89$ , (e.g., Miniard, Bhatla, and Rose 1990)

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

Attitude-Toward-the-Brand (Ab2),  $\alpha = .91$ , (e.g., Loken and Ward 1990)

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

**Table 2**

**The Effect of Webpage Background on Ad Effectiveness**

Variable	F-Ratio	Sig. level	MEANS		
			Black	Colored	Animated
<b>Multivariate test</b> (Wilk's lambda)	2.496	.045			
<b>Univariate tests</b>					
Aad1	3.474	.035	5.525*	5.158	4.650
Ab1	3.987	.022	4.911*	4.700	4.200
<b>Multivariate test</b> (Wilk's lambda)	1.939	.106			
<b>Univariate tests</b>					
Aad1	3.474	.035	5.525*	5.158	4.650
Ab2	2.219	.115	4.711	4.611	4.200
<b>Multivariate test</b> (Wilk's lambda)	2.099	.083			
<b>Univariate tests</b>					
Aad2	1.662	.196	5.411	5.033	4.744
Ab1	3.987	.022	4.911*	4.700	4.200
<b>Multivariate test</b> (Wilk's lambda)	1.406	.234			
<b>Univariate tests</b>					
Aad2	1.662	.196	5.411	5.033	4.744
Ab2	2.219	.115	4.711	4.611	4.200

\* In post-hoc comparisons (Tukey HSD) these means associated with the black background were found to be statistically different from those associated with the animated background. None of the means associated with the black background were statistically different from those associated with the colored background.

## Appendix

### Reviewers' Statements Relevant to Scale Standardization

*Journal of Advertising* #RR95-62 Reviewer A

*“ . . . 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.”*

*Journal of Advertising* #95-70 Reviewer D

*“Given that Aad is often conceptualized as a person’s global evaluation of the ad, just as Ao or Abrand 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 Aad as an attitude, one can simply adopt this standard attitude measurement, just as is done for Ao. There is no need for the Churchill paradigm in this situation. It is an irrelevant benchmark.”*

*Journal of Advertising* #94-28 Reviewer C

*“Given the assumption that there exists a domain of scale items that are representative of a construct, measures of Aad 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.”*

***Journal of the Academy of Marketing Science*** #0402-5-1 Reviewer #2

*“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 Aad?”*