Combating scale proliferation
Received (in revised form) 24th January, 2003

Gordon C. Bruner II
is an associate professor, Department of Marketing, Southern Illinois University, Illinois, USA. He is Director of the Office of Scale Research and the lead author of the Marketing Scales handbook series (AMA). His research has been published in Journal of Marketing, Journal of the Academy of Marketing Science, Journal of Retailing, Journal of Marketing Education and many other journals and conference proceedings.

Abstract: It is not unusual for new scales to be created in scholarly marketing research with little concern for their relationship with previous measures of the same construct and little justification provided for their development. This practice is challenged and it is argued that this is one of the things that is preventing marketing from progressing as a science. An empirical example is provided to illustrate how this lack of concern can have negative consequences on a stream of research. Increased emphasis on scale standardisation, equivalency and justification are suggested as major components of the solution to the problem of scale proliferation. A scale selection and usage paradigm is offered to guide researchers in reaching these triple goals.

Usage of multi-item scales has become common over the last couple of decades in conducting scholarly market research. What also is clear, however, is that for most constructs there is little agreement about which pre-existing scale to use or whether it is just as good to create a new scale for the purpose. A study of attitude-toward-the-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. 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 eg attitude-toward-the-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 the typically provided evidence of a scale’s reliability and validity is important but insufficient reasoning for use of a new scale in a study. The potential negative effects that result from the practice of scale proliferation will be described and illustrated. As a response to this problem it is argued that standardisation, equivalency and justification should take on greater importance in scale selection and usage paradigm. Finally, the paper concludes by offering a guide that typical users of scales may follow as they select, use and/or create scales.

BACKGROUND

Standardisation

With regard to psychometric scales, standardisation 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 that principle is violated if scientists can disagree about the measure. The implication is that a field's evolution from art to science is impeded as long as researchers can measure constructs any way they see fit because there is no way to say one scale is better than another.

There are several reasons for standardising usage. First, some degree of standardisation 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. Similarly, no single use of a scale is likely to examine all of the facets of generalisation that should be addressed by any measure for which wide application is desired. Thirdly, a proper 'final' step in scale construction is development of norms. 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. Thus, multiple studies across conditions and populations should be conducted over time utilising the same scale in order to achieve the important psychometric goals of measure validation, generalisability 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. 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 per cent of scales were found to have been used just once. Similarly, for the period of 1960–1969 in psychology, 63 per cent of scales were used just once. Although there has been no known broad examination of marketing scale usage, a study of consumer scales estimated that 79 per cent of those published during the 1980s were used just once.

Beyond these general reasons for the importance of standardisation, 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 attitude towards-the-ad have been used, the conclusions drawn about the significant relationships of one version were different from those based on the other version(s). This diversity of operationalisation is a possible reason why significant variation in findings was found across studies in a major meta-analysis of advert attitudes.

Equivalency

While standardisation has to do with use of the same scale, equivalency suggests that it is possible to use alternative measures of the same construct and yet arrive at the same conclusion. Analyses can be conducted to help determine the degree of equivalency between scales that are supposed to measure the same construct. Measures can range from being parallel to congenic to arbitrary. If it is established that a set of scales are parallel then it can be safely assumed that use of any of them will lead to the same conclusion. This is also possible with scales that are congenic if
transformation rules are known and are applied. That assumption cannot be made with arbitrary scales since they are merely purported to measure the same construct.

When justified, covariance structure analysis can be a powerful means of assessing equivalency. Another possibility is similarity analysis. As argued by Steenkering, the advantage of similarity analysis over covariance structure analysis is that it can be used with small samples and is more amenable to the exploration of structure. There are even ways of doing some preliminary equating of measures when few or no data from respondents are available. This amounts to employing item response theory and "collateral information" such as expert opinions or theory.

As important as equivalency is argued to be when multiple scales are used to measure the same construct in the same culture, it becomes even more important in cross-cultural research. A fundamental issue is to determine what extent findings represent "real" differences in cultures or, instead, are merely measurement artifacts. Douglas and Craig have discussed the various types of equivalency that should be examined when conducting "national" studies, and Mulloy has illustrated empirical techniques for testing them.

Even though it is possible to examine equivalency as a means of justifying the use of new measures, it is rarely done. Instead, it has not been unusual for new measures of a construct to be developed even when one or more established scales exist. Calling this a "forsworn state of affairs," Jacoby criticized this practice and gave examples apparent in the late 1970s. The practice has continued, however, as illustrated in Table 1. These appear to be some of the "worst" cases of scale proliferation in scholarly marketing research that have occurred since Jacoby made his observations.

How is this practice defended? It has been 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. This is based on the domain sampling model in which "parallel" scales can be developed and lead to essentially the same conclusions when used in empirical research. Although conceptually appealing, the equivalency of scales 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 unlikely that such a process has been used much in marketing. Instead, items have been constructed which, at best, produce "alternate form" scales with unknown

<table>
<thead>
<tr>
<th>Construct</th>
<th>Sample of different scales</th>
</tr>
</thead>
<tbody>
<tr>
<td>Attitude toward the ad</td>
<td>Homer and Palmer, Mitchell and O'Keefe, Chandon and Freestone, Luoma et al., Dubray and Highley, Barril and Siver, O'Guin and Schan, Yeung and Task, Peterson, Wison and Brown, Gottfried and Politzer, Rich, Letters, Ridgeway and Heterogeneity, Kettler and Schermerhorn, Dodds, Monroe and Grewal, Fry et al., Shim and Garvey, Rothergum, Peters and Wedel, Bhargh, Leigh and Weltz, Dziez and Gharib, Yihua and Hurr, Yafft and Adner, Gremler et al., Pustotki and Lord, Kopp and Jaccard, Peterson and Farnham.</td>
</tr>
<tr>
<td>Attitude toward the brand</td>
<td></td>
</tr>
<tr>
<td>Innovation</td>
<td></td>
</tr>
<tr>
<td>Job satisfaction</td>
<td></td>
</tr>
<tr>
<td>Organisational commitment</td>
<td></td>
</tr>
<tr>
<td>Price consciousness</td>
<td></td>
</tr>
<tr>
<td>Product quality</td>
<td>Petroshas and McNair, Kilbourne, Paolino and Ridley, Brown, O'Guin and Schan, Yeung and Task, Peterson, Wison and Brown, Gottfried and Politzer, Rich, Letters, Ridgeway and Heterogeneity, Kettler and Schermerhorn, Dodds, Monroe and Grewal, Fry et al., Shim and Garvey, Rothergum, Peters and Wedel, Bhargh, Leigh and Weltz, Dziez and Gharib, Yihua and Hurr, Yafft and Adner, Gremler et al., Pustotki and Lord, Kopp and Jaccard, Peterson and Farnham.</td>
</tr>
<tr>
<td>Purchase intention</td>
<td></td>
</tr>
<tr>
<td>Size clarity</td>
<td></td>
</tr>
<tr>
<td>Shopping enjoyment</td>
<td></td>
</tr>
</tbody>
</table>

This list was limited to three different scales for ten different constructs. A more detailed list is available upon request from the author.
statistical equivalence. This introduces various forms of error, notably systematic
differences, that are not covered by the
domain sampling model.30

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. Scale creation is expected to at least provide evidence of unidimensionality and internal consistency if not convergent and discriminant validity as well.33,34
Providing this sort of evidence is also important for measures that are modified in some way. Lessons could be learned from the field of psychology where 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.32

Even when an established scale is used it is helpful to provide some limited justification. 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.

The lack of clear answers to these questions decreases the confidence that can be placed 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. Indeed, some reasons for not using pre-existing scales are quite 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 lack sufficient validity or claims of validity made for them are in doubt.33,34

Putting exceptions such as these to one side, the rule is that researchers should use previously developed measures unless they can explain why it is not possible or appropriate.33,34

AN EMPERICAL ILLUSTRATION

As stated above, standardisation, equivalency and justification may be considered unnecessary depending upon one's understanding of the domain sampling model. Specifically, some researchers may believe that if a scale appears on the surface to tap into the same semantic domain as a pre-existing scale then it is pointless to provide confirming evidence of that judgment.

The shortcoming of this assumption is demonstrated here as simply as possible by showing the potential consequences of using different scales to measure the same construct. Researchers may not be fully aware of the potential inconsistency in conclusions that can result because it is rare for them to use two or more multi-item scales in a study to measure the same construct. Instead, what is presented here is more akin to what would happen if the same relationships were being examined in studies conducted independently by researchers utilising different measures.

An experiment was designed to compare the effects of different visual backgrounds on a variety of responses to an advertisement at a website. Specifically, it was anticipated that an advertisement in the context of an elaborate, dynamic background would be associated with poorer attitude toward the ad and attitude toward the brand than plainer backgrounds.35 The sample was composed of 90 college students.
randomly assigned to three treatments. A little over half of the sample was male (59 per cent) and 54 per cent were between the ages of 20 and 24.

Included in the questionnaire were two mutually exclusive sets of items for measurement of attitude-toward-the-ad (Aad) and attitude-toward-the-brand (Ab). One set of items (Aad1) was derived from Mitchell and Olson22 and has been used since the 1980s by many others. The second scale (Aad2) was taken from Muehling, Stoltman and Mishra23 (1990), a popular nucleus of items that has been augmented with other items in various studies. Regarding attitude-toward-the-brand, version one has been used by Mintzard, Bhatta and Rose24 and others while the second version of attitude-toward-the-brand has been used by Loken and Ward25 among others. Items for all scales as well as their respective reliabilities are shown in the Appendix.

The effect of webpage background on attitude-toward-the-ad and attitude-toward-the-brand was tested by means of a MANOVA as well as complementary ANOVAs. The results are presented in Table 2. A total of four MANOVAs were run, each using a different pair of attitude-toward-the-ad and attitude-toward-the-brand scales. This was intended to simulate the varying conclusions that could be drawn in four independent studies of the same relationships if different pairs of scales were used.

The results indicated that Aad1 and Ab1 were significantly influenced by the independent variable (different webpage backgrounds) whereas Aad2 and Ab2 were not. This, in turn, led to the varying results of the MANOVAs. Depending upon the set of scales used, the conclusions could range from full support for the hypothesised relationships, to partial support, to no support. More generally, this illustration shows how the use of different scales to measure the same relationships can lead to different conclusions even though the scales may superficially appear to be acceptable alternative measures of the focal constructs.

DISCUSSION

The primary call made here is for standardisation. The ideal way to accomplish this is to have one theoretical conceptualisation for primary constructs eg attitude-toward-the-ad, attitude-toward-the-brand, etc.; as well as a preferred measure for each of them. That is, however, an unrealistic goal for the near future. Alternative views are likely to coexist for some time with a variety of measures related to each conceptualisation.

In the meantime, the procedure outlined below is offered as a guide for the selection and development of scales. With the spirit of standardisation in full view, the emphasis is on effort to identify and use previously developed scales. One reason researchers may have overt-developed scales in the past is simply because they were unaware that acceptable scales were already available. This has become a less defensible excuse since the 1990s, however, due to the ability to easily conduct computerised searches of databases in libraries and the web.

Further, books totally devoted to the topic have been published within marketing26,27 as well as related fields.28-30

Recommended scale selection and 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 possible due to 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 available scales handbooks.

4. If alternative scales are 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 scales most accepted by the different schools of thought.

   Ultimately, 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, it should be sufficiently described in any scholarly paper written about the research. 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, internal consistency is overwhelmingly the most prevalent type reported in marketing; (iii) indications of any substantive 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 of 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 a measure using widely accepted procedures should be developed:

   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 is 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 (d) (above). Chief among the added information would be, (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 and reported; (iii) if such information cannot be included
with the published article due to length restrictions, an appendix should be made available directly from the author or from the publishing journal’s website.

One of the tenets of the proposed paradigm is that development of new scales should be conducted only if necessary and the process ought to carry with it certain validation and reporting responsibilities. Understanding of the domain sampling model should include the realisation that the judgment of researchers will differ in the definition of a construct (the semantic domain) as well as what constitutes an adequate selection of items from that domain. The burden of evidence is on developers of new scales; not reviewers or readers; authors must explain why existing scales are inadequate and cannot be used.

The benefits of following this guide can accrue to individual researchers as well as to the discipline as a whole. For researchers themselves the benefits come in the form of efficiency and effectiveness. They are more efficient because it probably saves less time to locate and use a previously published scale of acceptable psychometric quality than it does to create a new one of similar quality. Researchers can be more effective because when they use the same measures as were employed in the work on which they are building then they can be more confident in their conclusions than if a totally new measure is used.

There are several benefits for the field as well. If a growing proportion of researchers follow these guidelines then the unwanted variance in findings should be decreased due to the reduction in the variety of scales being used to measure the same thing (method variance). As the variance in findings is reduced then the quantity and quality of principles upon which marketing operates will be improved. This in turn is one of the
requirements for moving the discipline down the road towards being a science.

It is acknowledged that typical users of scales 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, the field would benefit from more journal articles that critically compare competing scales that have been developed for popularly measured constructs. Good examples are the work done on innovativeness, which consumes values and market orientation. In some cases, the reviews might only judge the relative superiority of a scale over others based upon a review of the evidence from secondary sources. Ideally, however, there will be many more studies conducted that gather primary data to simultaneously test alternative measures with analyses that are rigorous enough to gauge the equivalency of the scales. The value of these articles is that other researchers can refer to them to make more informed choices among available measures. Such articles could also note areas for future research such as when all available scales have serious limitations and the development of better measures is justified.

Short of having the results of such reviews and comparisons as guides, 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 or similar 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. Given this, it is recommended that literature reviews, meta-analyses and syntheses of findings across studies be conducted more cautiously. If the scales used in a set of 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.

SUMMARY AND CONCLUSIONS

The purpose of this paper has been to describe a problem that is affecting marketing as it did its parent disciplines: scale proliferation. Not only is this practice wasteful but it is probably injecting large amounts of undesirable variance into results and conclusions. The primary solution advocated here is standardization, use of the same scale by all who want to contribute to a stream of research. Short of that, use of another scale may be necessary and/or desirable at times but only if justification is provided. Part of the justification that is called for when an alternative scale is used rather than an established one is evidence of equivalency. Thus, these three qualities of scale usage, standardization, justification and equivalency deserve much greater attention than given in the past if marketing is to advance beyond an art to become a science.

References

27. Contact the author for the quotes from reviews at top marketing journals that illustrate the view representative of those who place blind faith in the domain sampling model.
30. Churchill (1979) op. cit.
34. Churchill (1979) op. cit.
40. Loken, B. and Wind, J. (1999) 'Alternative approaches to understanding the determinants of

Combating scale proliferation

81 Churchill (1979) op. cit.
82 Gefen and Anderson (1988) op. cit.

**APPENDIX**

**Characteristics of scale examined**

<table>
<thead>
<tr>
<th>Attitude-toward-the-ad (Aad1), $\alpha = 0.91$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 good/bad</td>
</tr>
<tr>
<td>2 like/dislike</td>
</tr>
<tr>
<td>3 irritating/not irritating</td>
</tr>
<tr>
<td>4 interesting/uninteresting</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Attitude-toward-the-brand (Abd2), $\alpha = 0.91$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 unappealing/appealing</td>
</tr>
<tr>
<td>2 unattractive/attractive</td>
</tr>
<tr>
<td>3 unpleasant/pleasant</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Attitude-toward-the-brand (Ab1), $\alpha = 0.89$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 dislike/like</td>
</tr>
<tr>
<td>2 unfavourable/favourable</td>
</tr>
<tr>
<td>3 negative/positive</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Attitude-toward-the-brand (Ab2), $\alpha = 0.91$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 bad/good</td>
</tr>
<tr>
<td>2 poor quality/high quality</td>
</tr>
<tr>
<td>3 unsatisfactory/satisfactory</td>
</tr>
</tbody>
</table>