ECONOMICS & Sociology

Walter Wymer, Helena Maria Baptista Alves, A Review of Scale Development Practices in Nonprofit Management and Marketing, Economics & Sociology, Vol. 5, No 2, 2012, pp. 143-151.

A REVIEW OF SCALE DEVELOPMENT PRACTICES IN NONPROFIT MANAGEMENT AND MARKETING

Walter Wymer
University of Lethbridge,
4401 University Drive,
Lethbridge, Alberta T1K 3M4
Canada
Tel.: (403) 329-2111
E-mail: walter.wymer@uleth.ca

Helena Maria Baptista Alves
University of Beira Interior,
Research Unit NECE, Portugal
strada do Sineiro, s/n
6200-209 Covilhã Portugal
Tel.: +351 275 319 600
E-mail:
helena.mb.alves@gmail.com

Received: July, 2012
1st Revision: August, 2012
Accepted: October, 2012

ABSTRACT. We describe a set of recommended practices for scale development research in nonprofit management and marketing. General process issues are described followed by recommendations for EFA and CFA components of scaling research. Implications for researchers, journal editors and reviewers are discussed.

JEL Classification: L31 ***Keywords***: Scale development, scaling procedures

Introduction

The fields of nonprofit management and nonprofit marketing have been emerging as their own disciplines. Although initial research into these areas generally involved applying concepts from commercial management and marketing research into the nonprofit context, this simple application into the nonprofit context proved disappointing. It became increasingly apparent to nonprofit management and marketing scholars that the operationalizations of commercial management/marketing variables (that is, their measures) were inappropriate for use in measuring these concepts in nonprofit sector settings). In some cases, the measures could be adapted by rewording scale items to make them meaningful in the nonprofit context. However, in other cases, substantial changes in a measure implicate the concept being measured. If a measure's dimensions differ or the scale items differ when a concept is measured, then either there are two underlying concepts or the original concept or its measure lack validity.

It is not difficult to understand the incentives that encouraged early nonprofit management and marketing researchers to base their work on concepts and their measures developed for commercial management and marketing phenomena. These researchers were trained in the commercial context. A body of research had already been established. There was a need for acceptance by the research community used to thinking within a commercial context.

The need for nonprofit scholars to develop appropriate conceptualizations of constructs and valid measures for those constructs is now salient in many areas. Indeed, the further development of the nonprofit management and marketing fields requires this. Our work, then, will make a contribution by examining the state of the field with respect to scale development work in the nonprofit management and marketing fields. Given this substantive need in the nonprofit research community, we believed it was timely and important to (1) identify scale development research in our areas, (2) assess its quality, and (3) recommended best practices. The contribution of this study is twofold. First, our work will facilitate the identification of scales for the nonprofit research community. Second, our work will provide a guide for future scale development research to be done with accordance with sound methodological procedures.

In reporting our study, we will first describe our process for identifying extent scale development research. Then we will describe commonalities among these studies. Afterwards, we will recommend best practices for future development research. Finally, we will conclude with recommendations for researchers, journal editors, and article reviewers. Our common goal is to encourage more high quality scale development research in our fields so that knowledge discovery can continue and our fields mature.

Content analysis procedure

A literature search was undertaken to identify studies published whose primary purpose was the development of new measures in the fields of nonprofit management and nonprofit marketing. Bekkers (2010) published a report listing journals in which articles on philanthropy are published. He identified 11 core journals which served as our initial pool of journals to search. A subsequent search involved using various keywords in the Google Scholar search engine. Additionally, we searched electronic databases for journal articles we may have missed in the preceding searchers. Finally, we searched dissertation abstract electronic databases in order to identify relevant theses and dissertations.

Using a procedure similar to Worthington and Whittaker (2006), a paid graduate student assistant conducted the searches previously described. We were looking for published studies on new scale development, studies which refined existing scales, and studies which examined the validity of existing scales. The graduate student assistant was instructed to include a study, if not sure of its appropriateness, and the investigators would examine these possible cases for inclusion later. The initial pool of 33 was reduced to 20 after the investigators analyzed each study.

We examined a number of characteristics of the studies. We were first interested in the general process used. How well were the constructs of interest conceptualized? How thorough was the initial item pool generation process? Were separate samples collected for each phase of the scale development program? Did the study include an exploratory factor analysis stage (EFA) followed by a confirmatory factor analysis (CFA) stage? For studies reporting EFA and CFA procedures, we examined a number of characteristics.

We had several observations from our analysis. Most salient among these was that the scale development work generally did not conform to standard scientific methods for developing psychometrically valid measures. There was a large range in processes used in these studies. Some were quite good, while others were less so. Rather than highlight deficiencies, we believe it is more useful to present some recommendations for our research community to improve the quality of measures we use in our research. If we are accurately defining and measuring our constructs, then this will improve the quality of our research and help to advance our field.

The process of scale development research

We will describe a recommended process which we believe will provide some quality control assurance for investigators conducting scale development research.

Conceptual development

The essential first step is to define the concepts for which one is seeking to develop a measure. It is common for concepts to be undefined, weakly defined, or incorrectly defined in prior research. Furthermore, it is common for journal articles to define the same concept differently. Sometimes in prior research, one has to infer a concept's definition based on how the authors operationalize it in their research design. Therefore, it is important for future research to be quite denotative in describing their concepts under investigation.

For the scale developer, defining a concept may be the result of first explaining how the concept has been define explicitly or implicitly in prior research. Then describe the construct fully, describing its dimensionality meaning. If a construct is multidimensional, then the dimensions will also have to be defined. If a concept is poorly defined, its measure will be necessarily weakened. It is also important to differentiate the concept from other similar concepts.

The conceptualization or theoretical development of a construct can be guided by prior research (include other disciplines), denotative descriptions from dictionaries, connotative descriptions from other published works, interviews with experts and practitioners, and the research scholar's own insights. Feedback from colleagues is useful in improving the precision of the conceptualization and the process may take several iterations.

Weak conceptual development is perhaps one key weakness in much scale development research. There are several reasons for this. As mention, inadequate prior research is a problem. Also, researchers may be overly influenced by taking a statistical approach to scale development. While statistical methods are quite important at refining the initial item pool and establishing convergent and discriminant validity, a measure is useless if the variable it has been design to measure is not defined adequately.

Initial item pool

Once the concept has been well defined, the initial item pool needs to be developed. This can be informed by prior research. The definitions that have been developed in the prior step can be particularly helpful in working with colleagues and other experts to base item generation on the concept definition.

An initially large and comprehensive list of potential scale items can be reduced and refined by having colleague, students, and practitioners evaluate items, suggest different items, identify redundant items, and match items with their target dimensions. Anders and Gerbing (1991) discuss an interesting technique for assessing the quality of the initial item pool by evaluating their substantive validities which will require a pilot test. We recommend that substantive validities of the initial item pool be pretested prior to the first development study (EFA).

One important issue pertaining to the scale items is the need for reverse scored or negatively worded items. Researchers used to be taught to include reverse scored items to reduce response set bias. The idea was that participants completing questionnaires would have to read items more carefully if they contain some negatively worded items. Therefore, there would be less likelihood of participants marking the same choice number continuously without reading the item statement. While there is logic to this process, it often creates a spurious effect in the FA. The negatively worded items tend to load on their own factor, thus

distorting the factor structure of the EFA. It considered a better practice to avoid negatively worded items (DeVellis, 2011).

We cannot over-emphasize the importance of the conceptual development and the initial item pool development. If these procedures are not rigorous, investigators are unlikely to get their study published in a major journal. This initial work is time consuming. However, compared to the remaining process, it is inexpensive and requires few resources.

Empirical development of scale

The general process is to follow an iterative process of performing EFA analyses on an item pool, refining the item pool, then repeating the process. Once the item pool is refined, the a CFA analysis is needed to support the measurement model and to provide evidence of convergent and discriminant validity. A carefully developed initial item pool will greatly benefit the empirical developmental work.

Sampling becomes a salient issue during the empirical work. First, the sample should be representative of the population of interest. A convenience sample of students is inappropriate if the investigators are developing a measure for the quality of life of the elderly, for example. Another issue pertains to the number of samples. It is recommended that each iteration of the process uses a new sample. It is inappropriate to perform on CFA analysis on a measurement model developed from data used in the EFA analysis. We witnessed one study in which the researchers took one sample, split the sample in two, and then performed and EFA analysis on one half and a CFA analysis on the other half. This would only be arguably suitable if the EFA resulted in no refinement to the scale item pool. However, this process would not be our recommendation.

Related to the sampling issue is reporting aspects of the data. Researchers need to report how they treated missing data. They also should report statistics on the level of normality of their data. Unfortunately, reviewers often fail to appreciate that many phenomena produce skewed data and that normal data distributions are not common. Finally, studies should report enough information, such as correlations and standard deviations, so that other researchers can recreate a matrix summary of the data for additional work without having access to the original data (Kline, 2010).

The ordering of efa and cfa in new scale development research

Researchers generally use CFA after their measure has been assessed using EFA and they want to understand if the factor structure from the EFA analysis fits the data from a new sample. While it is possible to develop a new scale using multiple CFAs, we recommend using EFA initially and CFA finally.

One problematic issue deals with the differences between using factor analysis (FA) and structural equation modeling (SEM). SEM allows measurement errors to correlate, whereas FA does not. This leads some to conclude that SEM is a superior technique. However, SEM is bias toward fewer items. The implication is that the final scale will contain relatively few items as the researchers respecify the SEM measurement models to bring their fit statistics near or below cutoff values (e.g., RMSEA < .05). While proponents of SEM argue that CFA using this approach removes common method bias and improves the measurement models, it leads to overly parsimonious measures that may insufficiently measure the conceptual domain of the construct. Thus, one is left with the dilemma of having items that share measurement error or having too few items to adequately measure the construct.

Our recommendation is for investigators to follow-up their EFA analysis of the factor structure of the scale items with an SEM to detect problematic levels of error correlations.

This process will provide the investigators some information to assess item retention, deletion, or modification decisions. Scale developers should keep in mind, however, that error correlations *between* different dimensions are more problematic than error correlations *within* a dimension. We next discuss specific issues pertaining to EFA and CFA.

EFA[1]

Researchers need to report the characteristics of their sample. Is it a purposeful sample of a target group or a convenience sample? Researchers need to justify the size of their sample. Is it of sufficient size to support the EFA? A good practice is to have a sufficient sample size so that the participant to item ratio is greater than 5 to 1, preferably closer to a ratio of 10 to 1 (Worthington & Whittaker, 2006).

Another issue related to sample size pertains to the adequacy of the correlation matrix to produce a factored solution. It is incumbent upon scholars to provide evidence for scale factorability. We, therefore, recommend researchers report and meet the following criteria. Report the Kaiser-Meyer-Olkin (KMO) statistic. KMO values of .60 or higher are recommended.

With respect to extraction methods in FA, we recommend common factor analysis. We do not recommend principal-components analysis (DeVillis, 2011). There are several techniques of FA. We recommend principal-axis factoring.

With respect to the rotation method, we recommend oblique rotations. If researchers know beforehand that the extracted factors do not correlate, then orthogonal rotation would be acceptable. Otherwise, using orthogonal rotation when the factors correlate tends to overestimate loadings. Using orthogonal rotation may lead to inappropriate item retention or deletion and this rotation method's factor structure may be more difficult to confirm in the CFA analysis.

Researcher need to report the criteria they used in deciding which factors to retain. Generally, eigenvalues greater than 1 are used initially. We recommend also using an examination of the scree plot. Factors with eigenvalues greater than 1 and to the left of the elbow in the scree plot are good candidates. However, the interpretability of the factor is a central issue. If researchers cannot interpret the factor, if the contained items do not make sense as a group, then the factor should not be retained. Researchers should take caution that factors having fewer than three items may create difficulties in a subsequent CFA analysis using SEM.

Researchers need to report the criteria they used in determining which items to delete or to retain. Loadings and cross-loadings are most commonly reported. We recommend deletions of items with low loadings on their primary factor and minimal cross-loadings. A cutoff of less than .40 for item loadings is recommended. Items with cross-loadings having less than a .20 difference from an item's highest factor loading should be considered for deletion. Finally, items with communalities less than .50 may also be candidates for deletion.

Researchers need to consider the overall scale length. Longer scales (scales having a greater number of items) tend to be more reliable. Longer scales also tend to better measure the conceptual domain of the construct. However, scales taking longer than 15 minutes can strain respondents' motivation to complete the scale. Scale length is no longer the substantive

[1] ed. EFA Exploratory Factor Analysis (EFA) is a statistical approach to determining the correlation among the variables in a dataset. This type of analysis provides a factor structure (a grouping of variables based on strong correlations). EFA is good for detecting "misfit" variables. In general, an EFA prepares the variables to be used for cleaner structural equation modeling. An EFA should always be conducted for new datasets. http://statwiki.kolobkreations.com/wiki/Exploratory_Factor_Analysis

issue it once was because of the inclusion of SEM CFA analytic tool. As mentioned previously, SEM is biased toward parsimonious scales having relatively few items.

CFA[2]

SEM has become the most commonly used means of assessing the measurement models in CFA. Therefore, we will discuss the issues pertaining to using this CFA approach in scaling procedures.

As in EFA, researchers need to describe the appropriateness and characteristics of their sample. Researchers need to defend the adequacy of their sample size and describe any criteria that guided their desired sample size. Sample size as a function of participants per parameter is recommended. We recommend a 10:1 ratio of participants to parameters as a best practices, but we view a ratio of not less than 5:1 as acceptable.

The level of missing data and how the researchers dealt with missing data should be reported. We recommend using forced choice questions in the measure to eliminate missing data. However, if there is missing data and it does not exceed five percent of the total, then SEM applications have useful algorithms for treating missing data.

Also related to the quality of data is the issue of outliers. Outliers can affect the data's multivariate normality and they can alter the outcome of the analysis. SEM applications commonly provide statistics for identifying outliers, such as Cook's Distance or Mahalanobis distance. We recommend that researchers use the SEM procedure to evaluate the pervasiveness and magnitude of outliers. It is important to recognize that not only can outliers bias CFA estimates, they may reduce the reliability of your measurement model. It is unlikely that subsequent studies will have a similar pattern of outliers and may have different results as a consequence. We do not have specific recommendations regarding outliers, but we caution researchers to give this issue attention and to increase their desired sample size 10 percent to allow outlier removals, if needed. Our experience has been that removing the most egregious outliers improves the reliability of the measure.

Overall model fit is traditionally tested by a chi-square statistic. A chi-square probability value greater than .05 indicates an acceptable model fit. Chi-square values should be reported with degrees of freedom, sample size, and p values. Chi-square is biased by sample size and, although it is customary to report this statistic, its importance has waned over time. Other fit indices are available which are now used as salient statistics to assess model fit against cutoff values.

Our recommendation is to report all fit indices. Researchers have different preferences and what is customary today may change in the future. Our current recommendation is to make the TLI and RMSEA fit indices the salient indices and to view a TLI of .95 or greater and an RMSEA of .05 or less as desirable optimal target cutoff values for determining a superior fit of the model to the data. Having given these recommendations, however, we believe that an RMSEA of .07 is acceptable if respecifying the model to achieve the .05 cutoff would require the deletion of scale items that are deemed important in adequately measuring the conceptual domain of the construct. We recommend reporting the RMSEA's 90 percent confidence range (lower and upper values). It is desirable if the upper value does not exceed .08.

Frequently, the measurement model does not demonstrate an initial good fit with the sample data. Subsequently, researchers then respecify the model in order to achieve an

[2] ed. Confirmatory Factor Analysis (CFA) is the next step after exploratory factor analysis to determine the factor structure of your dataset. In the EFA we explore the factor structure (how the variables relate and group based on inter-variable correlations); in the CFA we confirm the factor structure we extracted in the EFA. http://statwiki.kolobkreations.com/wiki/Confirmatory_Factor_Analysis

acceptable fit. Hence, CFA can become an exploratory approach to refining the model. If the changes made to achieve a good fit are minor, then this process can be considered confirmatory. An example of this may be to add a covariance parameter between two indicator errors of the same dimension. Major changes or an inability to achieve an acceptable level of fit may indicate additional item changes are required, necessitating an additional CFA study. If our preceding recommendations are followed, the researchers should not be markedly surprised by the CFA results. That is, we recommended that the EFA studies be accompanied with pilot CFA examinations of the data in order to detect issues related to measurement error correlations. For example, in such a pilot CFA analysis, it is possible to identify a scale item that is highly correlated with the error term of another item, resulting in a reduce level of fit. Researchers can then use this pilot CFA analysis to inform their item retention/deletion decisions.

We did not find the use of item parceling in studies we identified in our investigation. In some situations, parceling can provide a means of retaining a sufficient number of scale items to adequately measure a construct's conceptual domain while enabling the model to become parsimonious enough to produce good fit indices. Parceling is acceptable for unidimensional measures and in which the data is not normally distributed. Researchers should keep this possibility in mind.

Research need to establish the convergent and discriminant validity of their measures. We recommend that the CFA questionnaire include alternative measures, perhaps even simple single item measures with high face validity, that can used as a comparison. We believe this process is still useful. However, because it is now common to use SEM for CFA studies, a statistical approach for supporting a scale's convergent and discriminant validity is now feasible and convenient. Fornell and Larcker (1981) recommended a statistical process that is derived from the SEM output estimates. Their approach has gained acceptance as more CFA studies are conducted using SEM instead of FA, which was traditionally used. We recommend that the Fornell and Larcker approach be used in conjunction with a comparison of alternative measures.

Conclusions

As a result of our having identified and examined prior scale development research in nonprofit management and marketing, the need for improve standards and greater consistency became apparent. Therefore, to make this contribution we have discussed and recommended best practices based on our understanding of current scale development and psychometric research. There was not a sufficient body of scale development research in the nonprofit management and marketing fields to conduct a trend analysis. Indeed as more researchers are acknowledging the need for the discipline to develop its own theory, constructs, and measures, more rigorous scale development research has emerged in recent years. Nevertheless, this research has not been sequentially developed nor has it concentrated on specific topical areas in greatest need of research inquiry.

With reference to our investigation, we found gross inconsistency across scale development research in our discipline. Most disturbing was the tendency to use a single sample to develop and validate a scale. Since, the reliability and construct validity of much of the prior research is suspect.

The continued development of our discipline is dependent upon the quality of the conceptual development of the constructs we wish to study. It is also dependent upon the quality of the scales we use to measure the variables in our research. We hope our work will encourage the research community to support this effort and work collectively to provide

incentives for scale development research and to adhere to quality standards such as those recommended in this paper.

Recommendations for journal editors and reviewers

We would be remiss if we did not discuss the implications of our work for the gatekeepers of our discipline--journal editors and reviewers. Our experience has been that journal editors and reviewers have generally created an incentive system against producing scale development research, establishing a barrier to knowledge advancement in our field. We acknowledge that this is not the intention. However, it appears to be the unintended outcome.

To be more specific, it is not uncommon to find a journal that explicitly states that it is not interested in scale development articles. Although this is astonishing from a scientific perspective, it appears that some journal editors feel that scale development articles are common and, therefore, not interesting. Our search found a relatively small number of scale development articles in nonprofit management and marketing and we believe this work should be encouraged and rewarded.

Scale development research is time consuming, involving multiple studies. Writing an article that reports the conceptual development of the construct, the procedures used, and provides evidence of validity requires a number of pages. It is common, however, for journals' submission guidelines to place page restrictions on submissions that limit the ability of researchers to report a sound scale development research program. Ironically, researchers required page limits preventing them from adequately reporting their studies creates an environment in which reviewers' evaluate the submissions negatively because of inadequate reporting, conceptual development, and defending procedures. Further compounding this dynamic is that reviewers have varying levels of understanding pertaining to scale development research. Reviewers may have different beliefs regarding how scale development research should be conducted, what criteria should be used, and so forth.

To ameliorate this problems and reverse this vicious cycle we make several recommendations. First, we recommend that journal editors and editorial review boards develop written standards to scale development articles. The standards should state what procedures are required and what reported in needed. Recommended practices, cutoff values, and criteria should be specified. To accompany these standards, reviewers should be provided with a rating scheme from which to evaluate submitted scale development research. This will improve consistency between reviewers and help reduce variation in reviewer knowledge and preferences.

Second, we recognize that reviewer editors have journal issue page limits from the publisher. Publisher restrictions create a dilemma for the journal editor. Journal editors want to provide a means of disseminating research findings to the research community and briefer article allow for a greater number of articles in a journal issue with a total page per issue limit. To deal with this dilemma, we recommend that journal's publish their standards and reviewer rating guides so that researchers understand expectations. This will improve consistency and quality in scale development research. We also recommend that scale development research submissions not be constrained by an arbitrary page limit. Once the submission has undergone reviews and revisions, then a final, edited version of the scale development study can be published in an issue. Hence, the scale development submission is fully reported, standards are met, and the final article conforms to page restrictions. The editor can add a statement to the article that it has been reviewed in its entirety and meets the journal's standards. Obviously, the journal editor must decide what gets reported and not report in the final printed article. However, the full version of the article can be made available in the "available first" online version of the article, now a common feature among journals.

References

Anderson, J. C. & Gerbing, D. W. (1991), Predicting the performance of measures in a confirmatory factor analysis with a pretest assessment of their substantive validities, *Journal of Applied Psychology* 76(5), pp. 732-740.

Bekkers, R. (2010), *Journal in philanthropic studies*, Unpublished report, available by contacting author at r.bekkers@vu.nl

Clark, L. A., & Watson, D. (1995), Constructing validity: Basic issues in objective scale development. *Psychological assessment, 7*(3), p. 309.

DeVellis, R. F. (2011), *Scale development: Theory and applications* (Vol. 26): Sage Publications, Inc.

Hinkin, T. R. (1995), A review of scale development practices in the study of organizations, *Journal of Management, 21*(5), 967-988.

Kline, R. B. (2010), *Principles and practice of structural equation modeling*: The Guilford Press.

Fornell, C., & Larcker, D. F. (1981), Evaluating structural equation models with unobservable variables and measurement error, *Journal of Marketing Research*, *48*, pp. 39–50.

Rossiter, J. R. (2002). The C-OAR-SE procedure for scale development in marketing, *International Journal of Research in Marketing*, *19*(4), pp. 305-335.

Worthington, R. L., & Whittaker, T. A. (2006). Scale development research. *The Counseling Psychologist, 34*(6), pp. 806-838.