Original Article

The impact of scale width on responses for multi-item, self-report measures

Received (in revised form): 5th October 2011

Reto Felix

is an Associate Professor of Marketing at the University of Monterrey, Mexico. He received his Master's in marketing and PhD in business administration from the University of St. Gallen, Switzerland. He has published in journals such as the *Journal of International Marketing*, the *Journal of Business and Industrial Marketing* and the *Journal of International Consumer Marketing*. Further, he has presented his research at conferences hosted by the Association for Consumer Research, the American Marketing Association, the Academy of Marketing Science and the Society for Marketing Advances.

ABSTRACT Participants in market research studies are exposed to measurement scales with differing numbers of scale points, ranging from as little as 3 response options to over 20. However, the question whether variations in scale width may influence responses and bias the results remains the subject of ongoing discussion in marketing theory and practice. In this article, the impact of scale width on marketing scale properties is investigated in the context of two important marketing measures, attitude toward the ad (Aad) and attitude toward the brand (Ab). The results suggest that although, as expected, the average number of scale points used by respondents increases with the number of response alternatives available, scale width does not influence important indicators such as means, standard deviations and skewness. It follows that scale means can be compared between scales with differing numbers of response categories, using the general conversion formula presented in this article.

Journal of Targeting, Measurement and Analysis for Marketing (2011) **19,** 153–164. doi:10.1057/jt.2011.16; published online 14 November 2011

Keywords: scale width; response options; measurement scales; information processing; psychometrics

INTRODUCTION

Measurement scales are fundamental for marketing theory and practice and have become an indispensable instrument in survey research. In fact, the constructs for which measures have been developed are so manifold that attempts have been made to systematically collect and structure the available scales.[1,2] Most of the scales used in consumer research are multi-item, self-report measures that ask respondents to choose one among several categories related to a statement or question. However, a concern that many researchers seem to share is if the format of a particular measurement scale influences the responses obtained from a study. The two most important variations in scale format are related to (i) scale length (that is, the number of items in the scale), and (ii) scale width (that is, the number of response categories within every item). A substantial amount of research has focused on the impact of scale length on scale reliability.[3,4] The specific interest in scale length seems understandable, given the theory-based prediction that increasing the number of scale items usually increases the scale's reliability.[5]

Correspondence: Reto Felix
Department of Business Administration, University of Monterrey, Morones Prieto 4500 Pte., 66238 San Pedro Garza García, Mexico

However, it is the second aspect, scale width, which is of cardinal interest for researchers because the number of response categories is inextricably related to the information processing capabilities of respondents. On the one hand, it may be argued that a scale with very few response alternatives is possibly limited in its ability to discriminate.[6] For example, a respondent with 'true' rating values of 1.3, 1.4, 2.2 and 2.4 would be forced to choose 1, 1, 2 and 2 on a four-item three-point scale with response categories ranging from 1 to 3. Thus, the non-linear transformation from the 'true' metric scale to an integer scale may cause substantial measurement error and finally corrupt the conclusions derived. On the other hand, a scale with too many response alternatives may go beyond the respondents limited powers of discrimination[7] and thus increase the tendency to apply a *status quo* heuristic by selecting the response category that was selected for previous items.[8] These changes in response style may thus bias mean levels of responses or the correlation between marketing constructs.[9]

If both too few and too many response alternatives potentially lead to biased measurements, then the question arises if it is possible to determine an optimal number of response categories in a scale. Several empirical studies[3,7,10–12] have tried to provide an answer to this question by investigating what number of response categories optimizes scale reliability and validity, but the results are mixed and often contradicting. Further, the focus on optimizing scale reliability and validity conceals the fact that the concern of many researchers may rather be the consistency of means, standard deviations and skewness, as long as the scale's reliability and validity are within commonly accepted boundaries. For example, it is certainly interesting to know that a scale's reliability, as measured with Cronbach's α, is, for example, around 0.8 for a 3-point scale, increases to 0.9 for a 7-point scale and then decreases to 0.8 for an 18-point scale. However, as long as α is within a commonly accepted range, these differences in scale reliability may be of limited practical relevance, and the focus of attention might rather be the question: Are important scale indicators such as the scale mean affected by a change in scale width?

It is the objective of this article to investigate the influence of scale width on relevant indicators such as means, standard deviation, skewness and the number of response alternatives actually used by respondents. The article adds to previous research by including both stimulus and subject-centered scales and by explicitly investigating differences for positively and negatively skewed as well as approximately normally distributed scales. Further, it presents a universal formula for transforming evaluations between measures with differing scale width that is more general than the formula introduced by Dawes.[12] The article is structured as follows: In the first section, previous studies on scale width are reviewed and discussed. In the following section, Study 1 introduces the transformation formula for evaluations on scales with differing scale widths, and tests the impact of scale width for stimulus-centered, positively and negatively skewed scales (attitude toward three different print ads). Next, Study 2 expands on Study 1 by replacing the external stimulus (advertisement) from Study 1 with a memory recall task related to three different brands and by adding a subject-centered, approximately normally distributed scale (materialism). Finally, the results of both studies are discussed and implications for marketing theory and practice are suggested.

PREVIOUS RESEARCH ON SCALE WIDTH

Theory suggests that scales with more response alternatives also may have greater reliability, because, as Churchill and Peter[3] explain, 'variance may increase with the number of scale points and greater variance generally increases reliability' (p. 366). However, a review of research on the relationship between scale width and the psychometric properties of a scale does not lead to consistent conclusions. Some authors suggest that there is no influence of scale width on reliability.[7,10,13–15] Other researchers find at least weak relationships between scale width and either reliability or validity. Churchill and Peter[3]

hypothesized that the number of items, number of dimensions, difficulty of items, use of reverse-scored items and number of scale points are positively related to reliability (coefficient alpha). Only two measure characteristics had some impact on reliability estimates – the number of items that explained 10 per cent of the variance in reliability across all studies in their meta-analysis, and the number of scale points that accounted for 5 per cent of the variance. In a second meta-analysis, Peter and Churchill[16] conclude that measure characteristics such as the number of items, the number of scale points and the difficulty of items are positively related with reliability but not related with convergent and discriminant validity and only marginally related with nomological validity. However, the incremental R^2 on reliability owing to measure characteristics was relatively small (0.05). In a computer simulation, Bandalos and Enders[17] find that reliability increases with the number of scale points, but that, depending on the underlying distributions, maximum gains are obtained with five or seven response categories, after which reliability values level off. Comparable findings, based on empirical studies, have been presented by Birkett,[18] Finn,[19] Green and Rao,[6] McKelvie,[20] and Preston and Colman.[21] These findings suggest that the optimal number of response categories is around seven and thus provide support for Miller's[22] magic number seven plus or minus two in the context of human information processing capacity. Further, the question whether increases in reliability may be achieved with more response categories may also depend on the total score variability. Specifically, Masters[23] suggests that if low total score variability is achieved with a small number of response categories, an increase in response categories may lead to increased reliability. However, if total score variability is high, reliability seems to be independent of the number of response categories. On the other hand, there are speculations that an increased number of response categories may not increase reliability, but actually decrease it. Chang[11] finds that after removing systematic method variance from the measurement model, the six-point scale used in his study had less reliability than a four-point scale.

Apart from the effect of scale width on reliability or validity, a very limited number of studies report the effect of scale width on scale means or variance. Finn[19] finds an influence of the number of response categories on the mean of the ratings but remains cautious in his conclusions by stating that 'mean ratings appeared to be proportional to the number of scale levels but there were not sufficient data to justify fitting a curve' (p. 262). Aiken[10] demonstrates a linear increase in means and a curvilinear increase in variance with more response categories and concludes that scale width influences mean scores and variances. Dawes[12] finds slightly lower means (after rescaling) on a 10-point scale, in comparison to 5-point or 7-point scales, and no influence of scale width on standard deviations. Finally, Contractor and Fox[24] report that sensitivity, defined as the ability to statistically differentiate mean ratings for different stimuli, did not increase with increasing scale width. Rather, scales with five or six response categories showed slightly higher levels of sensitivity than wider scales with seven, nine or ten response categories. In conclusion, although theory suggests that scale width may have some influence on scale scores and the psychometric properties of a scale, empirical support for this influence is not very persuasive, and more research is needed to either dismiss or support previous findings from the literature.

STUDY 1

Study 1 examined if important scale properties are affected by a change in scale width for an attitude toward the ad scale. The study was conducted in the context of a rating task by asking participants to indicate their attitude toward three different print ads on a four-item, self-report scale. The ads for Study 1 were selected in a pretest that is described in the next section.

Pretest

Previous studies on scale width occasionally report skewness, but typically do not control for it *a priori*. So far, there is no theoretical

framework or empirical data available that consider a possible moderating influence of skewness on the relationship between scale width and scores such as means or standard deviations. In order to test two common scenarios, that is, a positively and a negatively skewed scale, the aim of the pretest for Study 1 was to assure that both a positively and a negatively rated advertisement would be included in the main study. Thirty-eight students from an undergraduate course at a private university in Northern Mexico were shown 10 print advertisements from magazines published in Spanish. The participants were asked to indicate their attitude toward the ads (Aad) on a four-item, seven-point differential scale, anchored by bad–good, low quality–high quality, unappealing–appealing and unpleasant–pleasant.[25] Items were averaged and the overall mean for the 10 advertisements was calculated. The ads selected for the main study were for the alcoholic beverage brand Martell ($m = 5.78$) and digital printing service Concepto ($m = 1.67$). A third ad from the watchmaker Tag Heuer was chosen as a control stimulus ($m = 4.25$). Because of the small number of participants in the pretest, reliability and validity were not calculated for the scale. However, these properties are reported for the main study.

Method

Four different versions of a questionnaire were assigned randomly to 103 undergraduate students who were enrolled at a private university in Northern Mexico. Students were asked to rate the three advertisements with the same attitude toward the ad scale used in the pretest. The scales for the first two advertisements (Martell and Concepto) were different for each version of the questionnaire: version 1 had three response options, version 2 had five response options, version 3 had seven response options and version 4 had nine response options. Odd-width scales were used because they presumably have higher reliabilities than even-width scales[4] and are preferable 'under circumstances in which the respondent can legitimately adopt a neutral position' (p. 402).[26] The width of the scale for the third ad (Tag Heuer) was held constant at seven points in order to verify the random assignment of the four questionnaire versions to the participants. After deleting two cases owing to missing data, 101 usable questionnaires were obtained.

Analysis

Rescaling. Because the Campbell and Keller[25] attitude toward the ad scale has seven response categories, the three-, five- and nine-point scales in the questionnaire were rescaled to seven points. For the rescaling, the following formula was used:

$$\mathrm{RV} = \mathrm{OV} \times \frac{\text{Scale Width of RS} - 1}{\text{Scale Width of OS} - 1} + \frac{\text{Scale Width of OS} - \text{Scale Width of RS}}{\text{Scale Width of OS} - 1},$$

where RV is the rescaled value, OV is the original value, RS is the width of the rescaled scale, OS is the width of the original scale. For example, in order to rescale the responses from version 2 of the questionnaire (five-point scale for ads one and two) to a seven-point format, the formula implies that a response of 1 on the five-point scale has to remain a 1 on the seven-point scale and that a response of 5 on the five-point scale has to be rescaled to a 7 on the seven-point scale. Further, to rescale, for example, a response of four on a five-point scale to its equivalent on a seven-point scale, the formula yields $4 \times 6/4 + (5 - 7)/4$, which equals 5.5. In order to illustrate the logic of the rescaling process, Table 1 shows the conversion of several example values from three-point, five-point and nine-point response values to a seven-point format. However, the formula works for any scale conversion in any direction. Dawes[12] did conversions from five- and seven-point scales to a ten-point scale, and the last two rows of Table 1 show that the formula is able to perform these specific conversions as well.

Dimensionality and Model Fit. Exploratory factor analysis (EFA) served to investigate the dimensionality of the Aad scales for Martell,

Table 1: Rescaling for different example scale values

From		*Calculus*	*To*	
Scale width	*Original value (examples)*		*Scale width*	*Rescaled value*
3	2	2×6/2+(3–7)/2	7	4
5	2	2×6/4+(5–7)/4	7	2.5
9	2	2×6/8+(9–7)/8	7	1.75
5	4	4×9/4+(5–10)/4	10	7.75
7	4	4×9/6+(7–10)/6	10	5.5

Table 2: Model fit for Aad

Advertisement	*Martell*	*Concepto*	*Tag Heuer*
Chi square	0.245	7.506	1.924
df	2	2	2
RMSEA	0.000	0.166	0.000
CFI	1.000	0.972	1.000
GFI	0.999	0.964	0.991
NFI	0.998	0.963	0.994

Concepto and Tag Heuer. All three scales loaded on one factor, with factor loadings ranging from 0.736 to 0.909. The fit statistics from the confirmatory factor analysis (CFA) are very satisfying (Table 2), except for the Root Mean Square Error of Approximation (RMSEA) for the Concepto ad that is above the recommended maximum level of 0.08 for a reasonable fit.[27] The Comparative Fit Index (CFI) is in all three cases above the 0.95 cut-off value suggested by Hu and Bentler,[28] and other indexes such as the Goodness of Fit Index (GFI) and the Normed Fit Index (NFI) are greater than the cut-off value of 0.90 for a good fit suggested by Kline.[29] Factor loadings from the CFA range from 0.618 to 0.882.

Reliability. Reliability was analyzed on an aggregate level across the four conditions. Cronbach's α of the measurement scales for the three advertisements (Martell, Concepto and Tag Heuer) was 0.83, 0.87 and 0.92, respectively, and thus exceeding Nunnally and Bernstein's[30] recommendation of 0.7 in each case.

Convergent and Discriminant Validity. Convergent validity was achieved, and all factor loadings from CFA were not only statistically significant,[31] but also substantial.[32] Specifically, all factor loadings exceeded 0.5. Discriminant validity was examined by calculating the average variance extracted (AVE) for each advertisement. AVE for Martell (AVE = 0.57), Concepto (AVE = 0.63) and Tag Heuer (AVE = 0.75) was greater than 0.5 and thus suggests good discriminant validity.[33]

Results and discussion

Mean Scores. The mean scores for the individual items and for the scale totals are shown in Table 3. If there was a relationship between scale width and scale means, then differences should be statistically significant for the Martell and the Concepto ad, but not for the Tag Heuer ad, because the latter served as a control by employing a seven-point scale for all four versions of the questionnaire. A one-way analysis of variance (ANOVA) suggests that the differences between the means shown in Table 3 are not statistically significant. Because it was intended to have advertisements in Study 1 that were positively as well as negatively rated, the distributions for most variables were either left-skewed (Martell and Tag Heuer) or right-skewed (Concepto), resulting in a violation of the assumption that data in a parametric ANOVA are normally distributed. Therefore, a comparison of mean values using the nonparametric Kruskal–Wallis one-way ANOVA by ranks was conducted. The Kruskal–Wallis one-way ANOVA by ranks is an extension of the Mann–Whitney *U* test to a design involving more than two independent samples and is based on the more relaxed assumption of ranked-ordered data. However, any transformation of interval/ratio data into ranks sacrifices information, which may explain why some researchers are reluctant to employ nonparametric tests, even when the assumptions

Table 3: Mean scores for items and scale totals (Study 1)

Variable	*Version 1 3-point (n=23)*	*Version 2 5-point (n=26)*	*Version 3 7-point (n=20)*	*Version 4 9-point (n=32)*	*Scale total for variable (n=101)*
Martell item 1	5.70	4.92	5.00	4.48	5.29
Martell item 2	6.09	5.56	6.05	5.62	5.79
Martell item 3	5.30	4.87	4.55	5.43	5.08
Martell item 4	6.09	5.33	5.70	5.52	5.64
Scale total Martell	5.79	5.17	5.33	5.51	5.45
Concepto item 1	2.57	3.13	2.65	3.04	2.88
Concepto item 2	2.43	3.02	2.65	2.80	2.75
Concepto item 3	1.91	2.56	2.30	2.29	2.27
Concepto item 4	3.48	3.31	2.70	2.88	3.09
Scale total Concepto	2.60	3.00	2.58	2.75	2.75
Tag Heuer item 1	6.13	5.77	5.95	6.03	5.97
Tag Heuer item 2	6.39	5.92	6.15	6.13	6.14
Tag Heuer item 3	6.26	5.88	5.75	5.97	5.97
Tag Heuer item 4	6.57	5.88	6.35	6.22	6.24
Scale total Tag Heuer	6.34	5.87	6.05	6.09	6.08

Note: All means are rescaled to a seven-point scale format.

Table 4: Standard deviations for items and scale totals (Study 1)

Variable	*Version 1 3-point*	*Version 2 5-point*	*Version 3 7-point*	*Version 4 9-point*	*Scale total for variable*
Martell item 1	1.77	1.70	1.17	1.47	1.56
Martell item 2	1.68	1.61	1.00	1.05	1.36
Martell item 3	1.77	1.95	1.73	1.65	1.78
Martell item 4	1.41	1.61	1.17	1.28	1.39
Scale total Martell	1.08	1.47	1.10	1.25	1.25
Concepto item 1	2.00	1.76	1.27	1.53	1.65
Concepto item 2	2.00	1.64	1.22	1.36	1.57
Concepto item 3	1.68	1.72	1.17	1.25	1.47
Concepto item 4	1.73	1.71	1.45	1.57	1.63
Scale total Concepto	1.38	1.53	1.05	1.34	1.34
Tag Heuer item 1	0.92	1.53	0.95	1.18	1.18
Tag Heuer item 2	0.94	1.47	1.04	0.98	1.12
Tag Heuer item 3	1.10	1.51	1.07	1.20	1.24
Tag Heuer item 4	0.59	1.63	0.81	0.91	1.09
Scale total Tag Heuer	0.78	1.45	0.80	0.96	1.04

of parametric tests are violated.[34] The Kruskal–Wallis one-way ANOVA by ranks suggests that the differences between the means in Table 3 are not statistically significant. In conclusion, the results from Study 1 suggest that scale width does have an impact on scale means.

Standard Deviations. There are changes in response style that may go undetected by changes in scale means. For example, if respondents select the endpoints of a scale (extreme response style) when scale width is either increased or decreased, and this increased use of endpoints would be distributed equally to the negative and positive endpoints of the scale, then this change in response style would not be reflected in a change of scale means. However, it should be reflected in a change of standard deviation. Thus, Table 4 shows the standard deviations for the four different versions of the Aad scale for each item and also for the scale totals. However, no specific pattern is recognizable from Table 4, and a Levene test for homogeneity of variance for each item separately as well as the scale totals suggests

Table 5: Skewness for items and scale totals (Study 1)

Variable	*Version 1 3-point*	*Version 2 5-point*	*Version 3 7-point*	*Version 4 9-point*	*Scale total for variable*
Martell item 1	−1.00	−0.40	−0.44	−0.86	−0.61
Martell item 2	−1.74	−1.33	−1.52	−0.91	−1.39
Martell item 3	−0.45	−0.54	−0.23	−0.72	−0.50
Martell item 4	−0.91	−0.81	−0.86	−0.54	−0.75
Scale total Martell	−1.07	−0.80	−0.70	−0.70	−0.84
Concepto item 1	0.93	0.36	0.40	0.02	0.45
Concepto item 2	1.10	0.43	0.19	0.14	0.56
Concepto item 3	1.74	0.95	0.21	0.57	1.01
Concepto item 4	−0.018	0.16	0.24	0.47	0.26
Scale total Concepto	0.84	0.65	0.10	0.26	0.56
Tag Heuer item 1	−0.66	−1.68	−3.10	−1.08	−1.40
Tag Heuer item 2	−1.27	−2.00	−0.95	−0.71	−1.62
Tag Heuer item 3	−1.25	−1.77	−0.88	−0.88	−1.30
Tag Heuer item 4	−1.00	−1.53	−1.42	−1.02	−2.01
Scale total Tag Heuer	−1.15	−2.12	−1.39	−0.86	−1.94

that there are no statistically significant differences between the scores.

Skewness. Table 5 shows that, overall, skewness for the three advertisements is as expected from the pretest. Martell and Tag Heuer are negatively skewed and thus obtained more positive than negative responses, whereas Concepto is positively skewed and thus received more negative than positive responses. However, there is no clear pattern related to the number of response categories, and the results suggest that scale width does not influence skewness of the scales.

Number of Unique Scale Points. A scale with more scale points offers more options to respondents and should be able to convey more information than a scale with fewer scale points.[26] However, this requires that respondents actually make use of the options offered in a scale with more response alternatives. Thus, the number of different scale points used across the four-item Aad scales was investigated for the four versions of the questionnaire. It seems plausible that for a nine-point scale, respondents may find it easier to use several different responses, for example, a seven on the first item and an eight on the second item, whereas on a three-point scale, a seven and an eight would both correspond to a three. Therefore, the average number of scale points used by respondents can be expected to be higher for scales with more response options.

Table 6: Number of unique scale points (Study 1)

Version	*Martell**	*Concepto***	*Tag Heuer*
1 (three-point)	1.78	1.78	1.83
2 (five-point)	2.12	1.81	1.81
3 (seven-point)	2.70	2.25	2.20
4 (nine-point)	2.44	2.28	1.91

*Differences are significant at $P<0.01$; **differences are significant at $P<0.05$

The number of unique scale points across the four-item Aad scales were counted casewise with the Excel command

```
=SUMPRODUCT((A1:A4<>"")
  /COUNTIF(A1:A4,A1:A4&"")).
```

The range A1:A4 in this example covers the cells for the four items of the Aad scale for a specific advertisement. If cases (respondents) were documented in the rows of a matrix such as Excel or SPSS, then for the next respondent the range would have to be adjusted to B1:B4. Table 6 shows the average number of scale points used by respondents for the four versions of the questionnaire. The averages of Martell and Concepto are rising monotonically, as expected, except for the difference between the seven- and the nine-point scale for Martell. An ANOVA confirms that these differences are statistically

significant ($P<0.05$). The differences for Tag Heuer (the control ad where only a seven-point scale format was used) do not show any clear pattern and are not statistically significant, which is also in line with the expectations. Thus, it may be argued that scales with more response options are able to convey more information than scales with fewer response options.

Discussion. The results of Study 1 suggest that scale width, that is, the number of response alternatives in a measurement scale, does not have any significant influence on important measures such as scale mean, standard deviation and skewness. Although it may be that differences did not show up because of measurement error, the psychometric properties of the three scales suggest otherwise. Specifically, unidimensionality was achieved for all scales, model fit indexes were overall satisfying, and all three scales proved to be reliable and showed good convergent and discriminant validity. On the other hand, theory suggests that wider scales are able to convey more information, and Study 1 found evidence supporting this theory. Specifically, the average number of unique scale points used by respondents showed a clear tendency to rise when scale width was increased. In order to test the results of Study 1 in a different context, Study 2 maintained the overall design of Study 1 but replaced attitude toward the ad with attitude toward the brand measures and added a subject-centered scale (materialism).

STUDY 2

In Study 1, participants were asked to rate an external stimulus (advertisement) based on visual perception. Study 2 substituted this external stimulus with a memory recall task triggered by showing the name and logotype for three different brands. No formal pretest was employed to select the three brands; however, informal talks with respondents belonging to the same segment as the main sample guided the selection process. Further, in distinguishing stimulus- and subject-centered scales,[35] Study 1 was limited to the use of stimulus-centered scales. Study 2 expands on Study 1 by adding a subject-centered scale to the questionnaire. Materialism was selected as a construct related to a subject-centered scale because the psychometric properties of the scale have been reported in several earlier studies and materialism has been shown to be of interest for marketing theory and of relevance for marketing practitioners.[36–38]

Method

Four different versions of a questionnaire were assigned randomly to 140 undergraduate students different from Study 1 who were enrolled at a private university in Northern Mexico. In the first section of the questionnaire, participants were exposed visually to the name and logotype of three well-known brands in Mexico: telecommunications company Telcel, computer and electronics producer Apple and clothing company Levi's. Participants were asked to rate their attitude toward the brand (Ab) on a four-item scale with the same anchors as the Aad scale used in Study 1. Consistent with Study 1, the scales for the first two brands (Telcel and Apple) were different for each version of the questionnaire and included three-, five-, seven- and nine-point scales for questionnaire versions 1–4, respectively. The width for the scale related to the third brand (Levi's) was held constant at seven points. Materialism was measured with the Richins[39] nine-item short form of the original Richins and Dawson[38] material values scale. The width for the materialism scale was manipulated the same way as for the Telcel and Apple Ab scales. After deleting 14 cases owing to missing data, 126 usable questionnaires were obtained.

Analysis

Rescaling. The conversion from the three-, five- and nine-point scales to a seven-point scale format was achieved by using the same formula as in Study 1.

Dimensionality and Model Fit. EFA revealed factor loadings ranging from 0.665 to 0.902 for the three Ab scales and confirmed unidimensionality of the scales. Factor loadings for the materialism scale revealed for item 4 a low factor loading of 0.445, whereas the loadings for the other

Table 7: Model fit for Aad

Advertisement	*Telcel*	*Apple*	*Levi's*
Chi square	1.449	0.009	0.030
df	1	1	1
RMSEA	0.060	0.000	0.000
CFI	0.996	1.000	1.000
GFI	0.994	1.000	1.000
NFI	0.989	1.000	1.000

eight items of the scale were in a range from 0.696 to 0.860. In contrast to the supposed three-factor structure of materialism with the well-documented dimensions of success, centrality and happiness,[38] all items loaded on one common factor. A closer inspection of the eigenvalues revealed that the eigenvalues for factors 2 and 3 were close to the default cut-off value of one for extracting factors in EFA. By overriding the default cut-off value and asking for a three-factor solution, the items loaded on the three dimensions of success, centrality and happiness as assumed by theory, except for item 4. This item is in fact the only reverse-worded item remaining in the Richins[39] nine-item short form, and problems with reverse-worded items in foreign (that is, non-US) contexts have been reported previously.[40] Because (a) materialism is a reflective rather than a formative measure, (b) for reflective measures 'construct validity is unchanged when a single indicator is removed, because all facets of a unidimensional construct should be adequately represented by the remaining indicators'[41] (p. 200), and (c) for multidimensional constructs (such as materialism), each dimension should be treated separately with its own set of observed variables,[42] item 4 was finally removed from the scale.

The fit statistics for Telcel and Apple were good, with exception of RMSEA. The fit indexes for Levi's were poor, especially RMSEA, which was much higher than the recommended maximum level of 0.08. A verification of the modification indexes for the three measurement models suggested a substantial error covariance between the errors for item 1 (bad–good) and item 2 (low quality–high quality). By allowing the errors for items 1 and 2 to covary (which comes at the expense of losing one degree of freedom), model fit for all three models improved considerably. It seems that for measurement models with relatively few indicator variables, and thus relatively few correlations and degrees of freedom, the omission of one single error covariance can decrease model fit substantially. Fit statistics for Telcel, Apple and Levi's after allowing errors 1 and 2 to covary are shown in Table 7.

Factor loadings from CFA for the three models without the error covariance estimation (that is, the models with two degrees of freedom) ranged from 0.657 to 0.910. Finally, the fit indexes for the three-factor, eight-item materialism scale were satisfying. Table 8 shows model fit indexes for the nine-item, one-factor; nine-item, three-factor; eight-item, one-factor; and eight-item, three-factor models. Factor loadings for the eight-item, three-factor materialism scale ranged from 0.486 to 0.851.

Reliability. Internal consistency of the measurement scales for the three attitude toward the brand scales were 0.77 (Telcel), 0.82 (Apple) and 0.90 (Levi's). For the nine-item materialism scale, reliability was 0.91 and for the materialism scale with item 4 removed, reliability increased slightly to 0.92.

Convergent and Discriminant Validity. Convergent validity was achieved for the three Ab scales as well as for the materialism scale with factor loadings being statistically significant and exceeding 0.5, except for item 3 of the materialism scale (0.49). AVE (a relatively stringent test of discriminant validity) was 0.47 for Telcel, 0.56 for Apple, 0.69 for Levi's and 0.44 for the materialism scale and thus above or at least close to the recommended cut-off value of 0.5.

Results and discussion

Mean Scores, Standard Deviations and Skewness. Among the three brands, Apple achieved the most positive ratings. The rescaled mean scores across the four scale versions were 5.33 for Telcel, 6.67 for Apple and 5.19 for Levi's. The mean for the seven-point materialism scale was 3.92 and skewness was close to 0, as predicted (0.045). As in Study 1, there was no clear pattern visible in the data, and the differences of means, standard deviations and

Table 8: Model fit for materialism

Model	*One factor Nine items*	*One factor Eight items*	*Three factors Nine items*	*Three factors Eight items*
Chi square	65.096	37.872	41.153	17.907
df	27	20	24	17
RMSEA	0.106	0.085	0.076	0.021
CFI	0.854	0.926	0.934	0.996
GFI	0.900	0.928	0.933	0.996
NFI	0.781	0.859	0.862	0.934

Table 9: Number of unique scale points (Study 2)

Version	*Telcel**	*Apple**	*Levi's*	*Materialism**
1 (three-point)	1.77	1.17	2.31	2.66
2 (five-point)	2.13	1.20	2.17	4.13
3 (seven-point)	2.27	1.61	2.27	4.00
4 (nine-point)	2.61	1.79	2.07	4.61

*Differences are significant at $P<0.001$.

skewness scores were not statistically significant, both for scale totals and individual items. Thus, tables are shown only for the number of unique scale points.

Number of Unique Scale Points. The average number of scale points for the four scales used in Study 2 is shown in Table 9. The averages of Telcel, Apple and the materialism scale were rising monotonically, except for the average of the seven-point version of the materialism scale, which was lower than the average of the five-point version (4.00 for the seven-point scale versus 4.13 for the five-point scale). The differences between these averages are statistically significant ($P<0.001$). Again, and in line with the expectations, the differences between the averages for the control scale (Levi's) are not statistically significant, and no specific response pattern can be seen.

Discussion. As in Study 1, the results of Study 2 suggest that scale mean, standard deviation and skewness are not influenced by changes in scale width. Reliability (internal consistency) was satisfying for all four measures used in Study 2. As in previous research,[40] reverse-worded items in the materialism scale proved to be problematic. Further, although model fit indexes (specifically, GFI, CFI and NFI) for the three Ab scales were overall quite good, RMSEA could only be improved satisfactorily by allowing the errors of item 1 and item 2 to covary. Finally, and in support of the findings in Study 1, the average number of unique scale points used by respondents increased when scale width was increased. For example, whereas on the four item, three-point Ab scale for Telcel respondents used only 1.77 scale points on average, on the nine-point scale this average rose to 2.61. For the control scale (Levi's), no systematic pattern in the average number of unique scale points could be detected, and the differences between the averages were not statistically significant, as expected.

GENERAL DISCUSSION

The results of the two studies suggest that, after rescaling, scale width, that is, the number of response categories, did not influence means, standard deviations and skewness of multi-item, self-report measures. These findings were robust for positively and negatively skewed distributions (Study 1) as well as for approximately normally distributed scales (Study 2). Further, the results apply to both stimulus-centered (Studies 1 and 2) and subject-centered (Study 2) scales. Although it might be argued that the absence of an effect on these three indicators is owing to a problem with the measurement instrument, there are two arguments that support the

findings: first, reliability as well as convergent and discriminant validity were satisfying in both studies, thus providing evidence for the adequacy of the measurement instruments. And second, scale width, as expected, influenced the number of unique scale points used by respondents. Specifically, the analysis showed that within the range of scale width used in both studies (three-point to nine-point scales), the average number of scale points used by respondents increased with the number of response alternatives available.

There are several implications of these findings for marketing academics and practitioners. From an information processing point of view, this article presents evidence that at least in the range from three-point to nine-point scales, respondents make use of the larger information capacity of wider scales. This finding supports the use of scales with more response categories. However, it is not clear if this tendency holds for wider scales, such as 11-point, 13-point or 15-point scales. Rather, it is probable that at some point, respondents are not able to differentiate between finer-graded response categories, and although, conceptually, information cannot be lost by increasing the number of response categories,[43] it may be expected that with increasing scale width, scales become too fine-graded for respondents and response errors such as *status quo* heuristics may cause respondents to use less instead of more unique scale points.[8,26] Future research thus should test the range where increasing the number of response alternatives leads to an increase in information conveyed by the scale.

From a more practical point of view, this article suggests that the selection of the optimal number of response categories is less problematic than previously hypothesized. Important measures such as means, standard deviations and skewness were not influenced by scale width, and these findings implicitly suggest that scale width does not influence important response styles such as acquiescence and extreme response style. This argument holds because although an influence of these response styles would not necessarily translate into changes in mean scores and skewness, it should affect the standard deviations of the scales. Although some caution should be advised, it seems that scale means, which are important indicators for decision making in marketing, can be compared between scales with differing numbers of response categories with a linear transformation, using the general conversion formula presented in this article. Researchers may thus be advised to pay more attention to content and predictive validity of scale items, instead of trying to optimize scale width. Further, in an ongoing attempt to simplify measurement instruments rather than complicating them, marketing practitioners frequently prefer the use of single-item measures instead of multi-item measures propagated by academics.[44] Recent research suggests that for constructs with a concrete object and attribute, such as the attitude toward the ad and attitude toward the brand construct, single-item measures are equally as valid as multiple-item measures.[45] Thus, future studies may investigate if the results presented in this article change when single-item scales are used.

REFERENCES

1 Bearden, W.O. and Netemeyer, R.G. (1999) *Handbook of Marketing Scales: Multi-item Measures for Marketing and Consumer Behavior Research*. Thousand Oaks, CA: Sage.

2 Bruner, G.C. (2009) Marketing scales database, http://www.marketingscales.com, accessed 15 July 2011.

3 Churchill Jr, G.A. and Peter, J.P. (1984) Research design effects on the reliability of rating scales: A meta-analysis. *Journal of Marketing Research* 21(4): 360–375.

4 Voss, K.E., Stem Jr, D.E. and Fotopoulus, S. (2000) A comment on the relationship between coefficient alpha and scale characteristics. *Marketing Letters* 11(2): 177–191.

5 Peter, J.P. (1979) Reliability: A review of psychometric basics and recent marketing practices. *Journal of Marketing Research* 16(1): 6–17.

6 Green, P.E. and Rao, V.R. (1970) Rating scales and information recovery-how many scales and response categories to use? *Journal of Marketing* 34(3): 33–39.

7 Jacoby, J. and Matell, M.S. (1971) Three-point Likert scales are good enough. *Journal of Marketing Research* 8(4): 495–500.

8 Weathers, D., Sharma, S. and Niedrich, R.W. (2005) The impact of the number of scale points, dispositional factors, and the status quo decision heuristic on scale reliability and response accuracy. *Journal of Business Research* 58(11): 1516–1524.

9 De Jong, M.G., Steenkamp, J.-B.E.M., Fox, J.-P. and Baumgartner, H. (2008) Using item response theory to measure

extreme response style in marketing research: A global investigation. *Journal of Marketing Research* 45(1): 104–115.
10 Aiken, L.R. (1983) Number of response categories and statistics on a teacher rating scale. *Educational and Psychological Measurement* 43(2): 397–401.
11 Chang, L. (1994) A psychometric evaluation of 4-point and 6-point Likert-type scales in relation to reliability and validity. *Applied Psychological Measurement* 18(3): 205–215.
12 Dawes, J. (2008) Do data characteristics change according to the number of scale points used? *International Journal of Market Research* 50(1): 61–77.
13 Bendig, A.W. (1954) Reliability and the number of rating scale categories. *Journal of Applied Psychology* 38(1): 38–40.
14 Komorita, S.S. and Graham, W.K. (1965) Number of scale points and the reliability of scales. *Educational and Psychological Measurement* 25(4): 987–995.
15 Matell, M.S. and Jacoby, J. (1971) Is there an optimal number of alternatives for Likert scale items? Study 1: Reliability and validity. *Educational and Psychological Measurement* 31(3): 657–674.
16 Peter, J.P. and Churchill Jr, G.A. (1986) Relationships among research design choices and psychometric properties of rating scales: A meta-analysis. *Journal of Marketing Research* 23(1): 1–10.
17 Bandalos, D.L. and Enders, C.K. (1996) The effects of nonnormality and number of response categories on reliability. *Applied Measurement in Education* 9(2): 151–160.
18 Birkett, N.J. (1986) *Selecting the Number of Response Categories for a Likert-type Scale*. Proceedings of the American Statistical Association. Washington DC: American Statistical Association, pp. 488–492.
19 Finn, R.H. (1972) Effects of some variations in rating scale characteristics on the means and reliabilities of ratings. *Educational and Psychological Measurement* 32(2): 255–265.
20 McKelvie, S.J. (1978) Graphic rating scales – How many categories? *British Journal of Psychology* 69(2): 185–202.
21 Preston, C.C. and Colman, A.M. (2000) Optimal number of response categories in rating scales: Reliability, validity, discriminating power, and respondent preferences. *Acta Psychologica* 104(1): 1–15.
22 Miller, G.A. (1956) The magical number seven, plus or minus two: Some limits on our capacity for processing information. *Psychological Review* 63(2): 81–97.
23 Masters, J.R. (1974) The relationship between number of response categories and reliability of Likert-type questionnaires. *Journal of Educational Measurement* 11(1): 49–53.
24 Contractor, S.H. and Fox, R.J. (2011) An investigation of the relationship between the number of response categories and scale sensitivity. *Journal of Targeting, Measurement and Analysis for Marketing* 19(1): 23–35.
25 Campbell, M.C. and Keller, K.L. (2003) Brand familiarity and advertising repetition effects. *Journal of Consumer Research* 30(2): 292–304.
26 Cox, E.P. (1980) The optimal number of response alternatives for a scale: A review. *Journal of Marketing Research* 17(4): 407–422.
27 Browne, M.W. and Cudeck, R. (1992) Alternative ways of assessing model fit. *Sociological Methods & Research* 21(2): 230–258.
28 Hu, L.-T. and Bentler, P.M. (1999) Cut-off criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling* 6(1): 1–55.
29 Kline, R.B. (1998) *Principles and Practices of Structural Equation Modeling*. New York: Guilford Press.
30 Nunnally, J.C. and Bernstein, I.H. (1994) *Psychometric Theory*. New York: McGraw-Hill.
31 Bagozzi, R.P., Yi, Y. and Philips, L.W. (1991) Assessing construct validity in organizational research. *Administrative Science Quarterly* 36(3): 431–458.
32 Steenkamp, J.-B.E.M. and van Trijp, H.C.M. (1991) The use of LISREL in validating marketing constructs. *International Journal of Research in Marketing* 8(4): 283–299.
33 Fornell, C. and Larcker, D.F. (1981) Evaluating structural equation models with unobservable variables and measurement error. *Journal of Marketing Research* 18(1): 39–50.
34 Sheskin, D.J. (1996) *Handbook of Parametric and Nonparametric Statistical Procedures*. Boca Raton, FL: CRC Press.
35 Torgerson, W.S. (1958) *Theory and Methods of Scaling*. Oxford, UK: Wiley.
36 Ahuvia, A.C. and Wong, N.Y. (1995) Materialism: origins and implications for personal well-being. In: F. Hansen (ed.) *European Advances in Consumer Research*. Provo, UT: Association for Consumer Research, pp. 172–178.
37 Burroughs, J.E. and Rindfleisch, A. (2002) Materialism and well-being: A conflicting values perspective. *Journal of Consumer Research* 29(3): 348–370.
38 Richins, M.L. and Dawson, S. (1992) A consumer values orientation for materialism and its measurement: Scale development and validation. *Journal of Consumer Research* 19(3): 303–316.
39 Richins, M.L. (2004) The material values scale: Measurement properties and development of a short form. *Journal of Consumer Research* 31(1): 209–219.
40 Wong, N., Rindfleisch, A. and Burroughs, J.E. (2003) Do reverse-worded items confound measures in cross-cultural consumer research? The case of the material values scale. *Journal of Consumer Research* 30(1): 72–91.
41 Jarvis, C.B., MacKenzie, S. and Podsakoff, P.M. (2003) A critical review of construct indicators and measurement model misspecification in marketing and consumer research. *Journal of Consumer Research* 30(2): 199–218.
42 Bollen, K. and Lennox, R. (1991) Conventional wisdom on measurement: A structural equation perspective. *Psychological Bulletin* 110(2): 305–314.
43 Garner, W.R. (1960) Rating scales, discriminability, and information transmission. *Psychological Review* 67(6): 343–352.
44 Gilmore, A. and McMullan, R. (2009) Scales in services marketing research: A critique and way forward. *European Journal of Marketing* 43(5/6): 640–651.
45 Bergkvist, L. and Rossiter, J.R. (2007) The predictive validity of multiple-item versus single-item measures of the same constructs. *Journal of Marketing Research* 44(2): 175–184.