More for Less? A Comparison of Single-item and Multi-item Measures

Marko Sarstedt
Ludwig-Maximilians-University of Munich

Marko Sarstedt/Petra Wilczynski

More for Less? A Comparison of Single-Item and Multi-Item Measures

Marko Sarstedt

Petra Wilczynski

■ **Schlüsselbegriffe**

Multi-Items; Reliabilität; Single-Items; Skalenentwicklung; Validität

■ **Keywords**

Multi-Items; reliability; single items; scale development; validity

Zusammenfassung

Der vorliegende Artikel vergleicht die Performanz von Single- und Multi-Item Messungen und erweitert die bis dato umfassendste Studie in diesem Forschungsgebiet von Bergkvist und Rossiter (2007), deren Ergebnisse aufgrund verletzter Annahmen zurückhaltend bewertet werden sollten. Auf Basis eines angemessenen statistischen Testverfahrens kann gezeigt werden, dass Multi-Items hinsichtlich Reliabilität und Kriteriumsvalidität signifikant bessere Ergebnisse liefern als Single-Items. Diese Erkenntnis stellt aktuelle Forschungsergebnisse in diesem Bereich in Frage.

Abstract

Despite their apparent practical advantages over multi-item measures, the use of single items is often regarded as a grave error, as they are believed to be unreliable and invalid. This article provides an evaluation of single-item measures regarding reliability and criterion validity and, thus, extends the most prominent study in this field by Bergkvist and Rossiter (2007) whose results should be considered with caution, due to a violation of testing assumptions. Using an appropriate testing procedure, we show that multi-item measures outperform single items to a significant degree, thus questioning recent research findings.

Autoren

Dr. Marko Sarstedt, MBR und Dipl.-Kffr. Petra Wilczynski, beide Ludwig-Maximilians-Universität München, Institut für Marktorientierte Unternehmensführung, Kaulbachstr. 45/I, 80539 München.

1. Introduction

While the utilization of multi items has been common practice in social science research for a long time, marketing researchers did not pay attention to the measurement of complex constructs and their validation until the 1970's. Instead, the application of simple single items was generally accepted[1]. Seminal works by Churchill (1979), Peter (1979), and Jacoby (1978) marked the start of a rethinking process by pointing out the deficits in earlier practices. These authors subsequently developed concepts for »better« approaches for measuring theoretical constructs by means of multiple items. These approaches are generally acknowledged as yielding more reliable and valid results, as they satisfy all psychometric requirements.

Conversely, single items are specifically conspicuous for their practical advantages like ease of application, brevity, and the low costs associated with surveys in which they are used[2]. Furthermore, long and complicated scales often result in a lack of understanding and mental fatigue, while single items usually promote higher response rates, as the questions can be easily and quickly answered. In addition, less effort is required for the development and analysis of research. Therefore, single-item scales are still prevalent among marketing practitioners, while many researchers claim that the use of single-item scales is a grave error. Thus, multi items are an implicit standard in academic research[3].

This gap between theoretical requirements and practical applications has led to considerable research in different fields over the last few years, which is reflected in the proliferation of various articles dealing with empirical comparisons of single and multi-item scales. Despite contrary beliefs, many of these studies conclude that single-item scales do not necessarily lag behind multi-item ones in terms of reliability or validity.

Wanous and Reichers (1996) introduced the first procedure to measure single-item reliability. Using the formula for correction for attenuation, the authors evaluated minimum reliability values of single-item measures of »overall job satisfaction«, employees' »perceived amount of participation in decision making«, and »desired amount of decision making«. Based on an empirical survey, they found an average minimum reliability for their single-item measures of .57 for the constructs under consideration[4]. Later, Wanous and Hudy (2001) replicated and extended this study in the context of »teaching effectiveness«, additionally using communalities derived from common factor analysis. Their estimate of .76 as the minimum level of single item reliability supports the previous study and shows that single items are not necessarily unreliable[5]. Using data for »belief in a just world« scales, Loo (2002) found that single items should only be used for extremely homogeneous constructs[6]. All these studies provide valuable insights into the performance of single-item measures. However, as they limit their objective to the evaluation of reliability, they do not comprehensively evaluate the psychometric properties of single-item measures, as reliable measures are not necessarily valid.

On the other hand, Ruekert and Churchill's (1984) analysis regarding »channel member satisfaction« only focused on single-item validity, using a correlation coefficient between the single and the respective multi-item scale as an estimate of convergent validity. They found correlations ranging from .58 to .68, thus showing that the measures exhibit convergent validity[7]. Likewise, Gardner et al. (1998) concentrated on construct validity; that is, convergent and discriminant validity when measuring the construct »focus of attention at work«. They concluded that »neither type of scale came out a clear »winner«.«[8].

The following studies provide a more comprehensive analysis as they consider both, reliability as well as validity. Wanous, Reichers, and Hudy (1997) published a meta-analysis of single-item measures of overall job satisfaction. They correlated each job satisfaction value with more complex multi-item measures, such as the well-known »Job Descriptive Index« and concluded that single items demonstrate an acceptable performance with regard to construct validity[9]. Their analyses' results of the estimated minimum level of reliability of the

[1] e.g., Homburg/Giering (1998), p. 113; Diamantopoulos/Winklhofer (2001), p. 270.

[2] e.g., Kwon/Trail (2005), p. 72; Bergkvist/Rossiter (2007), p. 176.

[3] Ryan/Buzas/Ramaswamy (1995), p. 11.

[4] Wanous/Reichers (1996), p. 633.

[5] Wanous/Hudy (2001), p. 369.

[6] Loo (2002), p. 73.

[7] Ruekert/Churchill (1984), p. 231.

[8] Gardner/Cummings/Dunham/Pierce (1998), p. 909.

[9] Wanous/Reichers/Hudy (1997), p. 249.

single-item measures ranged from .45 to .69, prompting the authors to conclude that »if the use of a single item is indicated, researchers may do so in the knowledge that they can be acceptable.«[10]. Nagy (2002) replicated this study for different facets of job satisfaction, adding behavioral items such as »fluctuation« and »productivity« to assess criterion validity, where single and multi-item measures showed comparable results[11]. Just like all studies mentioned before, Nagy's (2002) analysis clearly lacks convincing statistical evidence, as the author fails to test for significant differences between single and multi-item measures' validities. Furthermore, it is questionable whether the results for the constructs under consideration, which stem from the field of organizational psychology, can simply be transferred to the field of marketing. Consequently, these studies provide only limited guidance in terms of a comprehensive evaluation of the different scales, especially in a statistical context.

In this stream of research, the most recent and comprehensive study by Bergkvist and Rossiter (2007) brought a breath of fresh air into the literature on construct measurement and introduced the discussion to the field of marketing. In their study, the authors compare the criterion validity of single and multi-item measures of two of the most widely employed constructs in advertising literature, namely »attitude toward the ad« and »attitude toward the brand«. Based on a review of arguments for and against multi and single-item measures, the authors present three hypotheses relating to the measures' criterion validity, which is assessed by comparing bivariate correlation coefficients and R^2 values derived from ordinary least squares regression. In each case, the authors had to reject their hypothesis, as the differences between the coefficients turned out to be non-significant, prompting the authors to conclude that »this result fails to support the classic psychometric argument [...] that multiple-item measures are more valid than single-item measures for all types of constructs.«[12]. A detailed review of Bergkvist and Rossiter's (2007) analytical procedure, however, reveals a problematic aspect which challenges the study's results. In their analysis, the authors used Fisher's z standardization to test for significant differences between the predictor and criterion variables' correlations and R^2 values. However, according to classic statistical literature, this approach is not applicable in this case, as the test should only be used when two independent correlations from two different samples are considered[13]. In our study, as well as in that of Bergkvist and Rossiter (2007), the correlations stem from a single sample and are therefore not independent. Thus, the underlying assumption of the Fisher z-test – that covariances between the different correlations are equal to zero – does not hold. However, in the application of this test, even the slightest violation of testing assumptions is highly problematic, making it difficult, and sometimes impossible, to interpret the results in a meaningful way[14]. Conversely, by taking covariances into account, even very small differences between correlations can be significant. Furthermore, the authors segmented the data into very small subsamples, with sizes ranging between n = 55 and n = 86, thus reducing the power of the test[15].

In the light of this state-of-research, there's ample need for additional elaborations of the performance of single-item measures. Consequently, the aim of this study is to extend previous research on the appropriateness of single-item scales, most notably the study by Bergkvist and Rossiter (2007). Following their approach, we initially apply regression analyses, using averaged manifest variable scores as input data but compare the results, allowing for correlations among the coefficients. By also (inappropriately) applying the Fisher z-test to our data, we show that the previous objection is rather severe, because this approach leads to divergent results, thus calling Bergkvist and Rossiter's (2007) conclusion into question.

As another extension of previous research, which is often restricted to simple descriptive analyses, we apply Partial Least Squares (PLS) path modeling on the data. This is done for two reasons: Firstly, by using the PLS algorithm, we drop the assumption of equal manifest variable score weightings, as implied by Bergkvist and Rossiter's (2007) study. Secondly, we test the criterion validity of single items in a context that mirrors real-world applications in a better way, as single items

[10] Wanous/Reichers/Hudy (1997), p. 250.

[11] Nagy (2002), p. 83.

[12] Bergkvist/Rossiter (2007), p. 182.

[13] e.g., Cohen/Cohen (1975), p. 53; Bobko (1995), pp. 55.

[14] Hartung (1989), p. 354.

[15] In addition, the questions used to measure one construct were almost repetitive; see Bergkvist/Rossiter (2007), p. 180.

have frequently been used in PLS path modeling studies[16].

Lastly, we contribute to existing literature by testing Bergkvist and Rossiter's (2007) assumption that their study's results can be generalized to other popular constructs[17]. More precisely, we evaluate the interrelation between customer satisfaction and customer loyalty, ranging among the most salient constructs in marketing research. Consequently, we do not limit our analysis to single aspects of the measures, but compare single and multi-item measures by means of both, criterion validity and reliability.

The remainder of the paper is organized as follows: Based on a review of the key benefits and limitations of single and multi-item measures, we first derive several hypotheses for our research. At this, we primarily revert to the rationale presented by Bergkvist and Rossiter (2007). We then present different approaches to measuring single-item reliability and validity. Based on an exemplary model which examines the effects of customer satisfaction on customer loyalty in the German banking sector, we test our hypotheses in subsequent analyses. We also contrast our findings with those of Bergkvist and Rossiter (2007), which reveals that their analysis procedure should be viewed critically due to a violation of the underlying assumptions, and conclude with a discussion of the study's limitations and an agenda for future research.

2. Multi-Item vs. Single-Item Measures

The extent to which an issue is raised by successive generations of researchers and practitioners is a subtle indicator of its importance. The debate on the benefits and limitations of single and multi-item measures is an issue that has been heatedly debated in a variety of disciplines. To gain an understanding of the state-of-research, we review key arguments for and against both measurement approaches and derive our hypotheses based on our elaborations. It should be noted, however, that all these arguments only refer to reflective measures, as classic psychometric performance criteria are not applicable in respect of formative constructs. In addition, it is not possible to use a formative indicator as a single item, since all items affecting the construct have to be used to ensure a high degree of validity.

The most commonly mentioned advantages of single-item measures refer to the simplicity and brevity of surveying respondents and the lower costs associated with this[18]. The construction of the scale is quick and easy, while multi-item scales require a multi-stage construction process[19]. Furthermore, single items reduce the demands on participants. The use of single items is generally associated with lower levels of mental fatigue which yields higher response rates, an increased number of completed questionnaires, and thus leads to a survey's greater efficiency[20]. According to Nagy (2002), another advantage of single-item measures is their high flexibility, because they can be easily adjusted to new situations[21]. For example, when a new construct dimension has to be surveyed, a new multi-item scale has to be constructed in a multi-level process, while single items can be easily adjusted to meet the research objective.

Conversely, multi items are usually preferred, as they have psychometric advantages, especially with regard to greater reliability and validity. Reliability is increased since the use of multiple indicators adjusts random error. In other words, according to classical test theory and the domain sampling model, items are assumed to measure with error. The combination of numerous items therefore averages out this random error, which results in an adjusted measurement value, i.e. the true value. Hence, with an increase in the number of items, measurement error decreases while measurement accuracy increases[22]. As a larger set of adequate indicators covers a larger number of distinct construct facets, multi items also offer higher construct validity. In addition, use of multi items can equilibrate irrelevant aspects that are inherent in every question and are caused by the influence of other constructs. Thus, the items finally measure only the construct that is to be measured[23]. Kwon

[16] e.g., Gray/Meister (2004), p. 830; Venkatesh/Agarwal (2006), p. 375; Festge/Schwaiger (2007), p. 198.

[17] Bergkvist/Rossiter (2007), p. 183.

[18] e.g., Wanous/Reichers/Hudy (1997), p. 250; Nagy (2002), p. 77; Kwon/Trail (2005), p. 72.

[19] Gardner/Cummings/Dunham/Pierce (1998), p. 900; Jagodzinski/Manabe (2005), pp. 8.

[20] Drolet/Morrison (2001), p. 196; Bergkvist/Rossiter (2007), p. 176.

[21] Nagy (2002), p. 79.

[22] e.g., Churchill (1979), p. 66; Peter (1979), p. 7.

[23] Jacoby (1978), p. 93; Churchill/Peter (1984), p. 367; Kwon/Trail (2005), p. 71.

and Trail (2005) maintain that theoretical constructs are usually continuous[24]. However, since it is not possible to measure theoretical constructs on continuous scales, multiple items provide a better approximation of the data. Likewise, researchers often point out that using multi-item measures can help to handle the problem of missing values, as multi items provide several values for each construct, which allow item nonresponse to be efficiently solved through the application of imputation methods[25]. Applying multi items also enables researchers to better segment the data. This increases variability, allowing for a more precise segmentation of observations[26]. In this context, Bergkvist and Rossiter (2007) conclude that higher variability also leads to multi-item measures' higher correlation with a criterion[27]. Consequently, multi-item measures are assumed to have higher criterion validity[28]. Based on this line of argumentation and in accordance with Bergkvist and Rossiter (2007), we offer the following hypotheses:

H1: *A multi-item predictor correlates higher with a multi-item criterion than a single-item predictor with a multi-item criterion.*

H2: *A multi-item predictor correlates higher with a multi-item criterion than a multi-item predictor with a single-item criterion.*

H3: *A multi-item predictor correlates higher with a multi-item criterion than a single-item predictor with a single-item criterion.*

Ryan, Buzas, and Ramaswamy (1995) suggest that a construct measured with multi items can also explain more variance of an external criterion and therefore predict a relevant outcome in more detail[29]. Thus, the multi-item predictor should have higher predictive relevance for the criterion, which leads to hypothesis H_4.

H4: *A multi-item predictor has more predictive relevance for a criterion than a single-item predictor.*

In summary, it can be concluded that single items generally provide practical advantages, while multi items are conspicuous for their psychometric and methodical benefits (Table 1).

3. Assessment of Reliability and Validity

Reliability is the extent to which a scale produces consistent results if repeated measurements are made. This reflects the degree to which a measurement model is free from random error[30]. The most widely used measures to determine a measure's reliability are Cronbach's alpha and composite reliability. However, because their computation is based upon the correlations between the construct's items, these measures cannot be computed in respect of single items.

Alternatively, Wanous and Reichers (1996) suggest employing the classic correction for attenuation formula:

$$\hat{r}_{xy} = \frac{r_{xy}}{\sqrt{r_{xx} \cdot r_{yy}}}, \tag{1}$$

where r_{xy} is the correlation between construct x and y, r_{xx} (r_{yy}) is the reliability of construct x (y), and $\hat{r}_{xy}$ is an estimate of what the true correlation between the constructs would be, if they were perfectly measured[31]. The formula is usually applied to compare two different constructs, but can likewise be used in situations where x and y represent a single and a multi-item measure of the same construct. Consequently, it can be assumed that $\hat{r}_{xy} = 1$, which results in

$$r_{xx} = \frac{r_{xy}^2}{r_{yy}}, \tag{2}$$

where r_{xx} is the reliability estimate of the single-item measure x.

Wanous and Hudy (2001) suggest using communalities derived from common factor analysis to estimate single-item reliabilities. The authors argue that a single item's reliability must either be higher than or equal to its communality, as reliability additionally includes an item's specific variance. Consequently, communality can be consid-

[24] Kwon/Trail (2005), p. 71.

[25] Sloan/Aaronson/Cappelleri/Fairclough/Varricchio/Clinical Significance Consensus Meeting Group (2002), p. 481.

[26] Churchill/Peter (1984), p. 366.

[27] Bergkvist/Rossiter (2007), p. 176.

[28] see also Ruekert/Churchill (1984), p. 232.

[29] Ryan/Buzas/Ramaswamy (1995), p. 13.

[30] Peter (1979), p. 6.

[31] Wanous/Reichers (1996), pp. 632.

		Single-item measures	Multi-item measures
Theoretical aspects	**Reliability**	- no adjustment of random error - assessing reliability is problematic	+ allows for random error adjustment + determination of reliability by means of internal consistency
	Validity	- lower construct validity: • does not account for all facets of a construct • influence of other constructs - decreased criterion validity - assessing validity is problematic	+ higher construct validity • different facets of a construct can be captured • elimination of other constructs' influence + increased criterion validity + validity measures based on item-to-item correlations
	Segmentation tasks	- Segmentation solely based on the modalities of a single variable	+ more precise segmentation possible
	Missing values	- cannot be resolved efficiently	+ imputation methods based on correlations between indicators of the same construct
	Response behavior	+ confounding effects unlikely	- confounding effects possible
	Appreciation in academic research	- very uncommon (publication problematic)	+ generally accepted
Practical aspects		**Single-item measures**	**Multi-item measures**
	Costs	+ lower costs associated with scale development, questioning, and data analysis	- higher costs associated with scale development, questioning, and data analysis
	Non-response	+ increased response rate + lower item nonresponse	- lower response rate - higher item nonresponse
	Scale development	+ generally easy	- complex scale development
	Mental fatigue	+ low level of mental fatigue: simple, fast, and comprehensible	- increased level of mental fatigue: long, exhausting, and tiring
	Flexibility	+ high	- new scale development necessary

Tab. 1: Benefits (+) and limitations (–) of single and multi-item measures

ered a conservative estimate of single-item reliability[32].

These approaches serve as the basis for comparing single-item reliabilities with standard Cronbach's alpha coefficients for multi-item measures.

The standard psychometric approach for comparing validities is to examine how well each

[32] Wanous/Hudy (2001), p. 363.

measure predicts a relevant outcome[33]. This type of validity is referred to as criterion validity and is, according to Aaker, Kumar, and Day (2007), the most important kind with regard to the decision making process[34]. By considering the importance of customer satisfaction as a primary driver of customer loyalty, criterion validity's importance for the present study becomes apparent.

We first perform regression analyses by regressing (averaged) manifest variable scores for customer satisfaction on (averaged) manifest variable scores for customer loyalty. In doing so, we follow the analytical procedure as carried out by Bergkvist and Rossiter (2007). Likewise, we compute regression models based on standardized factor scores which yield results that are similar to the ones presented here[35].

Therefore, we calculate these in respect of different models:

- Customer satisfaction (multi items) → customer loyalty (multi items) [M_1]
- Customer satisfaction (single item) → customer loyalty (multi items) [M_2]
- Customer satisfaction (multi items) → customer loyalty (single item) [M_3]
- Customer satisfaction (single item) → customer loyalty (single item) [M_4]

The resulting standardized β-coefficients equal the constructs' correlations and can serve as a basis for testing hypotheses 1–3. Instead of using Fisher z-tests, we test differences between the correlation coefficients of the scales for significance by using a paired-sample test by Ferguson (1971), thus allowing for correlations between the coefficients:

$$\mathrm{t} = \frac{(r_{M_i} - r_{M_j})\sqrt{(N-3)\cdot(1+\mathrm{r}_{\mathrm{M}_{i\Delta j}})}}{\sqrt{2\cdot(1 - r^2_{M_i} - r^2_{M_j} - r^2_{M_{i\Delta j}} + 2\cdot r_{M_i}\cdot r_{M_j}\cdot r_{M_{i\Delta j}})}}\,. \tag{3}$$

In this formula, r_{Mi} (r_{Mj}) depicts the correlation between the constructs in model i (j), and $r_{Mi\Delta j}$ describes the correlations between the constructs that are not common to either model[36]. For example, when comparing M_1 and M_2, $r_{Mi\Delta j}$ depicts the correlation between the multi-item and the single-item measure of customer satisfaction. The test statistic follows a t-distribution with *N-3* degrees of freedom[37].

In the next step, we use PLS path modeling on the data to test the last hypothesis and to verify H_1 and H_2 in a more concise analytical way, as we drop the assumption of equal manifest variable score weightings. Originally developed by Wold (1974), PLS path modeling depicts an estimation method for path models, which serves as an alternative to the conventional covariance structure analysis by Jöreskog (1970)[38]. This approach has established itself over the last few years, especially in the management information systems discipline, as it is generally preferred when the requirements of interval scaled data and multivariate normality cannot be met. The latter is true of our study, as P-P-plots and Kolmogorov-Smirnov test results indicate that the data are non-normal[39].

Customer satisfaction's influence on customer loyalty is assessed across models M_1 and M_2, M_1 and M_3 as well as M_3 and M_4 by observing the path coefficients in the structural model. For the evaluation of path coefficients, t-values are calculated by using the bootstrapping procedure[40]. Multi-group comparison procedures may be applied to test model-specific coefficients for significant differences. In PLS path modeling, multi-group comparison is a rather new research field, which has only experienced ongoing development since Chin's (2000) introduction of the first approach. Other approaches apply permutation procedures or interpret group effects as caused by a categorical moderator variable[41]. However, as these procedures rely on independent samples, we modify Chin's parametric approach[42], reconciling it with paired samples by calculating the following test statistic, which follows a t-distribution with *K*-1 degrees of freedom:

$$t = \frac{\overline{z^*}\cdot\sqrt{K}}{\sqrt{\frac{1}{K-1}\cdot\sum_{k=1}^{K}\left(z_k^* - \overline{z^*}\right)^2}}\,, \tag{4}$$

[33] Bergkvist/Rossiter (2007), p. 173.

[34] Aaker/Kumar/Day (2007), p. 307.

[35] Bergkvist and Rossiter (2007) do not give a clear indication regarding their use of either mean manifest variable or factor scores. However, according to Rossiter's (2002, p. 310) C-OAR-SE article, mean values should be used to compute scores in reflective scales.

[36] Ferguson (1971), pp. 171.

[37] Applications of this approach can be found in Cohen/Cohen (1975), p. 53, Bobko (1995), p. 56, and Kwon/Trail (2005), p. 77.

[38] Henseler/Ringle/Sinkovics (2009).

[39] Bagozzi/Yi (1994), pp. 19; Chin/Newsted (1999), pp. 313.

[40] Henseler/Ringle/Sinkovics (2009).

[41] Chin (2003); Henseler/Fassott (2009).

[42] Chin (2000).

Item 1	Overall, how satisfied are you with your bank?	[1 = completely dissatisfied; 7 = completely satisfied 1 - 2 - 3 - 4 - 5 - 6 - 7
Item 2	Reconsidering all your experiences with your bank, how well did your bank meet your expectations?	[1 = did not meet my expectations; 7 = exceeded my expectations] 1 - 2 - 3 - 4 - 5 - 6 - 7
Item 3	When thinking of your ideal bank, how well does your bank compare?	[1 = very far away 7 = very close to my ideal] 1 - 2 - 3 - 4 - 5 - 6 - 7

Tab. 2: Operationalization of the construct customer satisfaction.

where z_k^* depicts the difference between the path coefficients that relate customer satisfaction and loyalty in models M_i and M_j in bootstrap run k (k $=1,\ldots,K$), i.e.. $z_k^* = y^{*k}_{Mi} - y^{*k}_{Mj}$. Thus, the average difference between path coefficients $\overline{z^*}$ in K bootstrap runs is given by

$$\overline{z^*} = \frac{1}{K}\sum_{k=1}^{K} z_k^* \ . \qquad (5)$$

The test is accurate if the populations, i.e. the bootstrapping values of the path coefficients, are normally distributed.

To test the last hypothesis, we revert to the cross-validated redundancy measure Q², which has been developed by Geisser (1974) and Stone (1974) to assess the predictive relevance of exogenous latent variables[43]. The redundancy measure can be computed using the blindfolding procedure, an iterative sample reuse technique that omits part of the data matrix during parameter estimation[44]. In each step, parameter estimates are used to calculate the omitted part. The residual variation of the estimates is then utilized to calculate Q². Values greater than zero imply that the exogenous construct has a predictive relevance for the endogenous construct, whereas values below zero reveal a lack of predictive relevance[45].

4. Research Design, Data, and Analysis

An exemplary model was used that measures the effects of customer satisfaction on customer loyalty. These constructs were chosen because they range among the most salient constructs in marketing research. Furthermore, customer satisfaction and loyalty are frequently treated as concrete constructs in terms of Rossiter's (2002) C-OAR-SE procedure and, hence, surveyed with single items[46]. In the survey, respondents were asked to rate their satisfaction with and loyalty to their private bank. Customer satisfaction was operationalized with three reflective indicators that had been frequently applied in past research[47]. Customer loyalty was measured by means of five reflective indicators that are well-known from empirical marketing studies and which exhibited a high degree of reliability and validity in past research[48]. All items were measured on a 7-point Likert scale (see Tables 2 and 3):

To compare the performance of the measurement approaches and to test the hypotheses, one item was chosen from each item set. Loo (2002) considers the results of a principal component analysis to identify the item with the highest factor loading as the single item of subsequent analyses[49]. However, this procedure does not appear to be reasonable. In common applications, a single item is not evaluated for goodness compared to other possible single items, but is rather created deductively as it is based on theoretical considerations. Thus, the single item has to be selected before any subsequent analysis. Consequently, we contacted eleven marketing experts from academia and practice to identify the salient item in each set. The experts consistently indicated that item 1 and item 6 should serve as the single-item measures.

Data were collected from N = 164 management undergraduate volunteers from the Ludwig-Maxi-

[43] Chin (1998), p. 317.

[44] Henseler/Ringle/Sinkovics (2009).

[45] Götz/Liehr-Gobbers (2004), p. 731.

[46] e.g., Anderson/Sullivan (1993); Cronin/Taylor (1992); Eberl (2009). Likewise, Reichheld (2003) introduced the »net promoter score« as a proxy for loyalty which has received great attention in marketing practice.

[47] e.g., Ryan/Buzas/Ramaswamy (1995); Fornell/Johnson/Anderson/Cha/Bryant (1996); Grønholdt/Martensen/Kristensen (2000).

[48] e.g., Zeithaml/Berry/Parasuraman (1996), p. 38.

[49] Loo (2002), p. 72.

Item 4	If required, I will use other services of my bank.	[1 = strongly disagree; 7 = strongly agree] 1 - 2 - 3 - 4 - 5 - 6 - 7
Item 5	I would recommend my bank to friends.	[1 = strongly disagree; 7 = strongly agree] 1 - 2 - 3 - 4 - 5 - 6 - 7
Item 6	I want to stay a customer of my bank for a long time.	[1 = strongly disagree; 7 = strongly agree] 1 - 2 - 3 - 4 - 5 - 6 - 7
Item 7	If I had to decide again, I would choose my bank.	[1 = strongly disagree; 7 = strongly agree] 1 - 2 - 3 - 4 - 5 - 6 - 7
Item 8	I can imagine purchasing other services of my bank.	[1 = strongly disagree; 7 = strongly agree] 1 - 2 - 3 - 4 - 5 - 6 - 7

Tab. 3: Operationalization of the construct customer loyalty.

Construct	Multi Items	Single Items	
	α_{xx}	r_{xx}	Communality
Customer satisfaction	.920	.960	.882
Customer loyalty	.905	.842	.762

Tab. 4: Results of the reliability analyses

milians University in Munich in 2007. These students' ages ranged from 20 to 54 (mean = 23.708; stddev = 3.955), and overall, 16 different banks were rated (for an overview of item means, standard deviations, and a correlation matrix, compare the Appendix). The software used for the analysis was SmartPLS 2.0 (M3), a comprehensive statistical software application for graphical latent variable path modeling[50]. For inside approximation, we applied the path-weighting scheme, which ensures well-predicted latent variable scores[51].

5. Summary of Results

In a first step, the single-item measures' reliability was evaluated according to the previously described approaches by Wanous and Reichers (1996) and Wanous and Hudy (2001). Table 4 illustrates the findings in respect of both presented reliability measures for single items, as well as the Cronbach's alpha coefficient (α_{xx}) in respect of the multi-item construct *x*.

The results show that multi items generally outperform single-item measures with regard to reliability. When following Wanous and Hudy's (2001) suggestion to consider communality as a conservative estimate of single-item reliability[52], single-item performance clearly lacks behind that of multi items. Only when using the mean of r_{xx} and the communality as a reference value, as also suggested by Wanous and Hudy (2001), this divergence is altogether less pronounced[53]. Nevertheless, single-item measures still perform acceptable with regard to reliability, as the estimates lie above the generally suggested threshold value of .70[54]. The results show that the assessment of single-item reliability is not clear-cut since the different methods' results vary. However, solely testing for reliability is not sufficient – reliable measures are not necessarily valid.

Table 5 provides an overview of the bivariate regression results to test for correlations between the constructs (hypotheses 1–3)[55].

[50] Ringle/Wende/Will (2005).

[51] Fornell/Cha (1994), p. 65.

[52] Wanous/Hudy (2001), p. 363.

[53] Wanous/Hudy (2001), p. 369.

[54] Nunnally (1978), p. 245.

[55] We also assessed convergent validity using a multi-trait-multi-method matrix. The results show that the multi and the single-item measure of the same construct correlate higher than two measures of distinct constructs. Likewise, we assessed the constructs' discriminant validity using the Fornell/Larcker-criterion (cp. the Appendix).

Model	Correlation / β_1
M_1	.799**
M_2	.758**
M_3	.762**
M_4	.716**

** Correlation significant at .01

Tab. 5: Results of the bivariate regression analyses

Models compared	Hypothesis tested	t-value	p-value
M_1 vs. M_2	H_1	2.499	.007
M_1 vs. M_3	H_2	2.257	.013
M_2 vs. M_4	H_3	1.645	.051
M_3 vs. M_4	H_3	2.602	.005

Tab. 6: Results of the correlation tests

Models compared	Hypothesis tested	z-value	p-value
M_1 vs. M_2	H_1	.936	.175
M_1 vs. M_3	H_2	.851	.197
M_2 vs. M_4	H_3	.826	.204
M_3 vs. M_4	H_3	.911	.181

Tab. 7: Results of the z-tests

The correlations between customer satisfaction and loyalty are uniformly high and differ only slightly. All correlations are significant at .01. A closer examination of the results reveals that in models that included multi items (M_1 – M_3), higher correlations can be observed. To test whether these differences are significant, we ran the paired-sample test by Ferguson (1971), using formula (3). In accordance with the formulation of hypotheses 1 – 3, we used one-tailed tests. Table 6 presents the results of these tests:

As shown, in the first two comparisons, the differences in the correlations are significant at a level of at least .013. Thus, by using multi-item scales, it is possible to achieve better results (i.e. higher correlations between the constructs). Consequently, we find support for hypotheses 1 and 2.

Hypothesis 3 cannot be tested directly, as it is not possible to test two paired correlations that are not correlated with a common element[56]. But since $\rho_{M1} > \rho_{M2}$, and $\rho_{M2} > \rho_{M4}$ in the population, and based on the transitivity principle, it can be concluded that $\rho_{M1} > \rho_{M4}$. Consequently, hypothesis 3 is supported by the data[57].

To compare our findings with those of Bergkvist and Rossiter (2007), we apply the Fisher z-test on our data. The results (see Table 7) show that the correlations between the models under consideration do not differ significantly. An evaluation of the hypotheses based on this procedure would have led to misleading conclusions as all of the hypotheses would have been rejected. The divergence between the tests is not surprising, bearing in mind that the correlations have built-in dependences if they are computed across the same individuals[58]. Consequently, Bergkvist and Rossiter's (2007) conclusion that the theoretical tests and empirical findings would be unchanged if commonly used

[56] Bobko (1995), p. 58.

[57] We also tested the correlations between the constructs on the basis of path modeling results, i.e. based on latent variable scores. A comparison by means of Ferguson's (1971) paired-samples test yielded equal results.

[58] Bobko (1995), pp. 55.

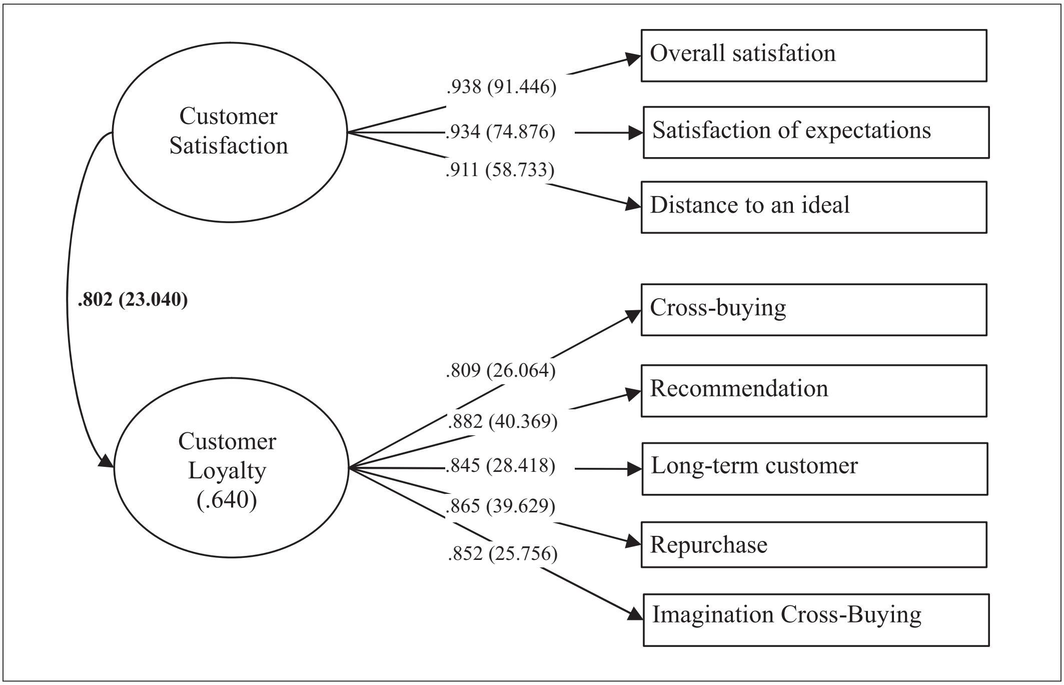

Fig. 1: Analysis results of model M_1

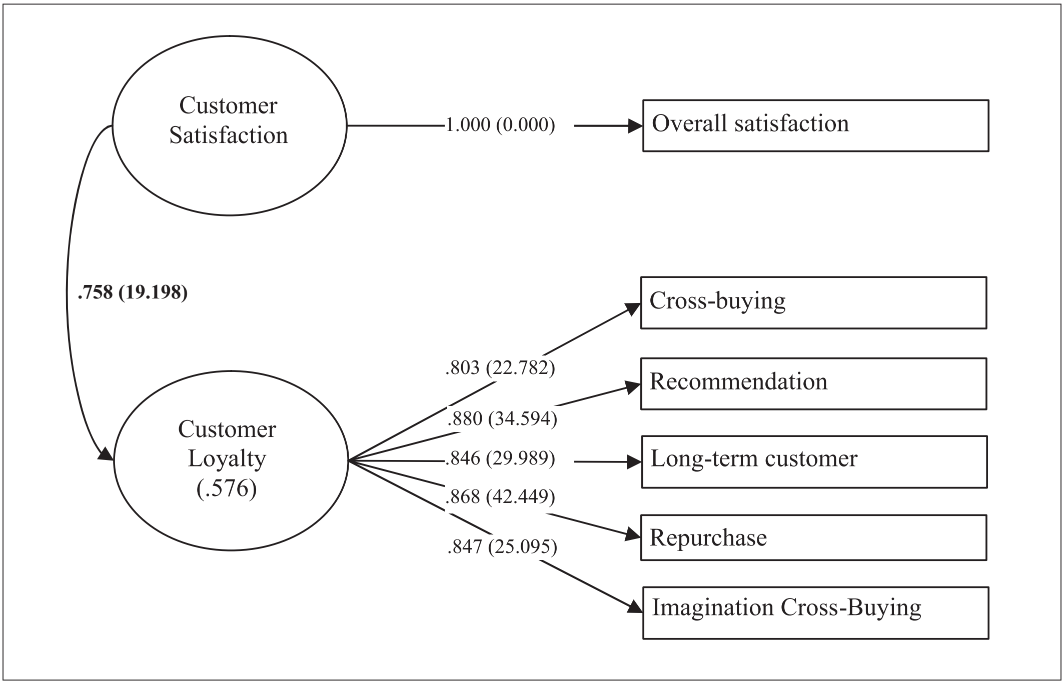

Fig. 2: Analysis results of model M_2

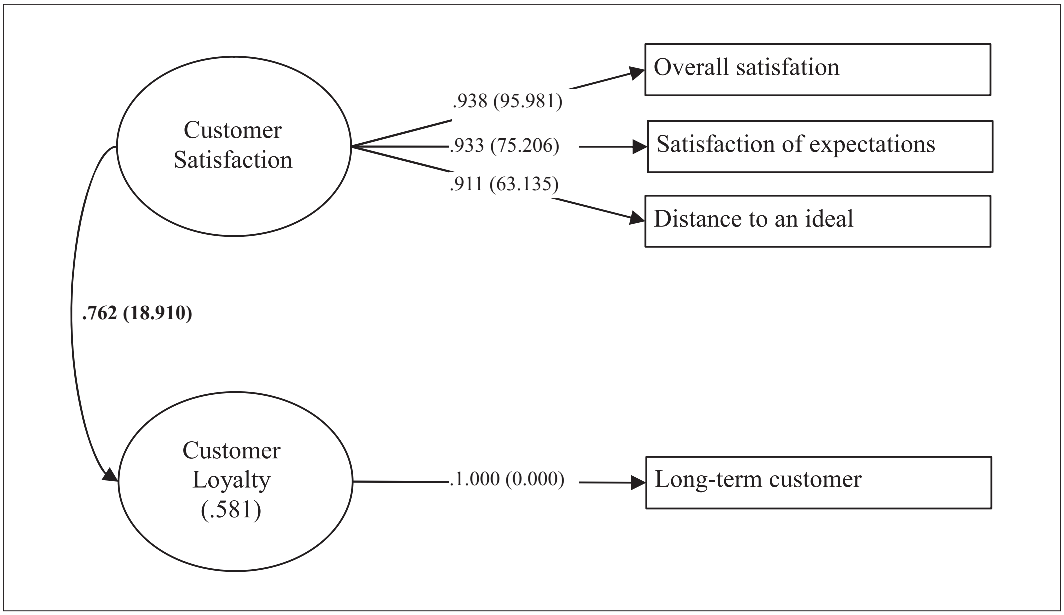

Fig. 3: Analysis results of model M_3

multi-item measures were substituted by good single-item measures should be considered with caution[59].

To find further support for H_1 and H_2 and to test whether a multi-item predictor has more predictive relevance for a criterion than a single-item predictor (H_4), we compared M_1 and M_2, M_1 and M_3 as well as M_3 and M_4 using PLS path modeling on the data. Figures 1–3 present the results of the first three models; t-values, based on bootstrapping results, are provided in brackets and R^2 values are shown in the icon of the endogenous construct.

The evaluation of the measurement models reveals that the constructs measured by multiple items are reliable (Customer satisfaction: composite reliability = .949, Cronbach's alpha = .920; Customer loyalty: composite reliability = .930, Cronbach's alpha = .905). All item loadings exceed the suggested value of .70, which implies that there is more shared variance between the construct and its measure than error variance[60]. High item reliability is also supported by the results of the bootstrapping procedure, which was carried out with 164 cases and 500 samples. The analysis reveals that all loadings are significant at $p < .01$.

When comparing the structural models, it is notable that the path coefficient of the multi-item model is slightly higher than that of the single-item models ($\gamma_1 = .802$ vs. $\gamma_2 = .758$ and $\gamma_3 = .762$). To evaluate whether these differences are significant, we calculate formula (4), using the bootstrapping results. The Kolmogorov-Smirnov test values as well as a visual inspection of P-P-plots indicated that the bootstraping samples do not deviate from the distributional assumption. The subsequent calculation of the test statistic shows that the coefficients differ significantly at p = .000 ($t_{1/2(499)}$ = 19.171 and $t_{1/3(499),}$ = 16.552), which furthermore supports hypotheses 1 and 2.

To test the hypothesis 4, we compute the cross-validated redundancy measure Q^2 and compare model M_1 with M_2 as well as M_3 with M_4. For M_1, the criterion lies at .451 and for M_2 at .404 whereas Q^2 takes a value of .576 for M_3 and .507 for M_4. Thus, all values are greater than zero and indicate that the predictor has predictive relevance for the criterion. In both cases, the value achieved when the predictor is measured with multi items is higher which suggests that the multi-item predictor has higher predictive relevance than the single-item

[59] Bergkvist/Rossiter (2007), p. 183.
[60] Hulland (1999), p. 198.

predictor and can therefore account for more criterion variance than the single-item predictor can. To test for differences between the model-specific predictive relevance measures Q^2, 157 blindfolding samples were drawn and tested for significance by means of paired t-tests[61]. With $p = .000$ ($t_{(156)} = 3.842$) for the comparison of model M_1 and M_2 and $p = .018$ ($t_{(156)} = 2.390$) for the comparison of M_3 and M_4, the test shows that a multi-item construct has a significantly higher predictive relevance than a single-item construct. Thus, we find support for hypothesis 4.

6. Discussion of the Results

This paper extends previous research on the appropriateness of single-item scales, most notably the most prominent study in this field by Bergkvist and Rossiter (2007), which has already been cited numerous times in top-tier academic journals. We argue that Bergkvist and Rossiter's (2007) study has to be considered with caution because the authors applied an inappropriate testing procedure. By disregarding built-in dependences among the correlations, the authors violate testing assumptions, which makes their interpretation of the results questionable. Consequently, we compare the performance of single and multi-item measures, applying an appropriate testing procedures, and using data from an own empirical survey. Contrasting our approach with that of Bergkvist and Rossiter (2007), we show that the initial objection is rather severe, thus challenging the authors' conclusion. In addition, by reverting to the PLS path modeling procedure, we use a more rigorous analytical approach which allows for differing weightings of manifest variables values. Moreover, we assess both the reliability and validity of single items.

Our analysis shows that from a strict psychometric point of view, single items significantly lag behind multi-item measures, thus confirming the classic psychometric line of argumentation in this respect. We obtained this result consistently for all methods of assessment, which fails to support Bergkvist and Rossiter's (2007) finding that single items and multi-item measures exhibit equal levels of criterion validity. This objection gains urgency, considering the nature of the constructs that were examined. Customer satisfaction and customer loyalty have become fundamental and well-documented constructs which are often surveyed in marketing studies and treated as concrete attributes in terms of Rossiter's (2002) C-OAR-SE procedure. This notion is shared by Bergkvist and Rossiter (2007), who speculate that their results should »generalize to other concrete attributes, such as beliefs or perceptions, intentions, and satisfaction«[62]. However, if single-item measures cannot match multi items even in the case of simple attributes such as satisfaction and loyalty, what about more abstract attributes such as corporate reputation, corporate social responsibility or organizational culture? In light of this, the use of single-item measures appears to be problematic for complex constructs, with regard to reliability and criterion validity. Consequently, Peter's objection[63] that most constructs are too complex to be measured effectively with a single item must not be prematurely discarded, as Bergkvist and Rossiter (2007) suggest when offering an almost concluding enumeration of potential (global) marketing constructs that can supposedly be measured with single items[64]. Apart from this, there remains the question how to identify the »best« element from an item set to serve as the single item[65]. Neither Rossiter (2002) nor Bergkvist and Rossiter (2007) give a clear indication on how this choice should be made. Loo (2002) selected the best item as the single item, basing this decision on factor loadings[66]. In our study, this indeed would have led to marginally improved single-item performance, but would not have altered the study's conclusion. Rammstedt et al. (2004) constructed distinct single items for each measured dimension. Their single items provided good results even in respect of the measurement of complex psychological constructs[67]. However, neither approach seems to be a good reflection of practical

[61] To establish a data basis for this test, we calculated the squared prediction error's proportion of the sum of squared errors using the mean for prediction for each blindfolding sample.

[62] Bergkvist/Rossiter (2007), p. 183.

[63] Peter (1979), p. 16.

[64] Bergkvist/Rossiter (2007), p. 183.

[65] Diamantopoulos (2005), p. 4.

[66] Loo (2002), p. 72.

[67] In fact, Rammstedt/Koch/Borg/Reitz (2004) obtained much better results than Loo (2002), although they measured far more complex constructs. This supports the notion that the applicability of single items largely depends on the construct under consideration and the chosen or constructed single item.

applications of single items. We believe that this question should not primarily be based on empirical considerations, as these procedures do not allow any findings to be generalized to other research situations. Consequently, in the present study, the single items were chosen from the results of expert interviews.

To summarize, our results and notions do not exactly give rise to optimism regarding the use of single-item measures on a broad basis. Lower levels of reliability and validity, and problems in the identification of an appropriate single item, on the one hand, face practical advantages on the other. Consequently, practitioners' modus operandi to favor single-item measures must be rejected. Having said this, one should bear in mind that single-item measures have not proven to be notoriously unreliable and invalid in absolute terms, as differences in reliability and validity are marginal (but significant). Despite these differences, but in light of their numerous practical advantages, researchers and practitioners could still argue that single items should be used more often in marketing research. Benefits in the form of higher response rates and lower costs could compensate for disadvantageous psychometric characteristics. However, researchers need a much clearer indication of the research situations in which the use of single-item measures is deemed appropriate.

Consequently, future research should aim at developing a systematic approach for the utilization of single items, taking into account considerations such as the complexity of the construct, the research objective or sample at hand. This includes the evaluation of other popular marketing constructs, such as corporate reputation or trust to test whether these can effectively be represented by a single item. A first research step would involve a thorough review of past research work to identify concrete constructs that can be represented by single items. Such systematization would allow a structured evaluation of the applicability of selected single-item measures. Future research should also address the limitations of this study: Our empirical study used a student sample, which may limit the generalizability of the results, as students have an above average educational background which is associated with, for example, higher flexibility, cognitive capability, creativity, and open-mindedness[68]. These abilities enable them to show higher levels of abstract thinking, which is required when answering single items as opposed to multi-item measures[69]. Consequently, the utilization of a representative sample could further reinforce the differences. In addition, future studies should compare single and multi-item scales by means of test-retest reliability which was not considered in our study. This research question is of particular interest as this type of reliability is strictly speaking the only stand-alone evaluation criterion of reliability, applicable to single-item scales. Researchers should also evaluate how single items will perform in more complex model set-ups with a greater number of constructs and path relationships. In this study, we used a very sparse model design with one exogenous and one endogenous latent variable. Consequently, the contrast between the application of PLS path modeling and simple ordinary least squares regressions is not very profound. In the methodological context, future research should also evaluate from which level of reliability, multi-item measures may be replaced by single items due to high levels of information redundancy. This could be achieved by means of simulation studies in which data constellations are systematically varied through the use of different model set-ups.

A richer understanding of the performance of single-item scales is necessary to foster their application in both marketing theory and practice – where appropriate. This should lead to a greater appreciation of this measurement approach to specific research situations in marketing journals.

[68] Fuchs/Sarstedt (2008).

[69] Sloan/Aaronson/Cappelleri/Fairclough/Varricchio/Clinical Significance Consensus Meeting Group (2002), p. 484.

Appendix

		Customer Satisfaction			Customer Loyalty				
		Item 1	**Item 2**	**Item 3**	**Item 4**	**Item 5**	**Item 6**	**Item 7**	**Item 8**
	Mean	4.970	4.762	4.396	4.628	4.610	4.497	4.72	4.402
	Stddev	1.463	1.452	1.395	1.547	1.652	1.726	1.791	1.619
Customer Satisfaction	**Item 1**	1.00							
	Item 2	.84	1.00						
	Item 3	.77	.76	1.00					
Customer Loyalty	**Item 4**	.58	.55	.60	1.00				
	Item 5	.72	.70	.71	.61	1.00			
	Item 6	.60	.59	.57	.56	.69	1.00		
	Item 7	.72	.68	.62	.55	.76	.72	1.00	
	Item 8	.61	.61	.61	.77	.64	.64	.62	1.00

Discriminant validity of the multi-item constructs:
AVE (customer satisfaction): .862
AVE (customer loyalty): .725
Squared correlation between the
constructs $.802^2 = .643$
→ According to the Fornell/Larcker-criterion, the measures show discriminant validity[70].

References

Aaker, David A./Kumar, V./Day, George S. (2007): Marketing Research. 9th ed. New York, NY 2007.

Anderson, Eugene W./Sullivan, Mary W. (1993): The antecedents and consequences of customer satisfaction for firms. In: Marketing Science, Vol. 12 (1993), No. 2, pp. 125–143.

Bagozzi, Richard P./Yi, Youjae (1994): Advanced topics in structural equation models. In: Bagozzi, Richard P. (Ed.): Advanced methods of marketing research. Cambridge 1994, pp. 1–51.

Bergkvist, Lars/Rossiter, John R. (2007): The predictive validity of multiple-item versus single-item measures of the same constructs. In: Journal of Marketing Research, Vol. 44 (2007), No. 2, pp. 175–184.

Bobko, Philip (1995): Correlation and regression analysis: Principles and applications for industrial/organizational psychology and management. New York, NY 1995.

Chin, Wynne W. (1998): The Partial Least Squares approach for structural equation modeling. In: Marcoulides, George A. (Ed.): Modern Methods for Business Research. London 1998, pp. 295–336.

Chin, Wynne W. (2000): Frequently asked questions – Partial Least Squares & PLS Graph. http://disc-nt.cba.uh.edu/chin/plsfaq/multigroup.htm, Accessed: December 09, 2008.

Chin, Wynne W. (2003): A permutation procedure for multi-group comparison of PLS Models. In: Vilares, Manuel J./Tenenhaus, Michel/Coelho, Pedro S./Esposito Vinzi, Vincenzo/Morineau, Alain (Eds.): PLS and related methods, PLS '03 International Symposium – Focus on customers. Lisbon 2003, pp. 33–43.

Chin, Wynne W./Newsted, Peter R. (1999): Structural equation modeling analysis with small samples using Partial Least Squares. In: Hoyle, Rick H. (Ed.): Statistical strategies for small sample research. Thousand Oaks, CA 1999, pp. 307–341.

Churchill, Gilbert A. Jr. (1979): A paradigm for developing better measures of marketing constructs. In: Journal of Marketing Research, Vol. 16 (1979), No. 1, pp. 64–73.

Churchill, Gilbert A. Jr./Peter, Jean-Paul (1984): Research design effect on the reliability of rating scales: A meta-analysis. In: Journal of Marketing Research, Vol. 21 (1984), No. 4, pp. 360–375.

Cohen, Jacob/Cohen, Patricia (1975): Applied multiple regression/correlation analysis for the behavioral sciences. New York, NY 1975.

Cronin, Joseph J./Taylor, Steven A. (1992): Measuring service quality: A reexamination and extension. In: Journal of Marketing, Vol. 56 (1992), No. 3, pp. 56–68.

Diamantopoulos, Adamantios/Winklhofer, Heidi M. (2001): Index construction with formative indicators: An alternative to scale development. In: Journal of Marketing Research, Vol. 38 (2001), No. 2, pp. 269–277.

Diamantopoulos, Adamantios (2005): The C-OAR-SE procedure for scale development in marketing: A comment. In: International Journal of Research in Marketing, Vol. 22 (2005), No. 1, pp. 1–9.

Drolet, Aimee L./Morrison, Donald G. (2001): Do we really need

[70] Fornell/Larcker (1981).

multiple-item measures in service research? In: Journal of Service Research, Vol. 3 (2001), No. 3, pp. 196–204.

Eberl, Markus (2009): An application of PLS in multi-group analysis: The need for differentiated corporate-level marketing in the mobile communication industry. In: Esposito Vinzi, Vincenzo/Chin, Wynne W./Henseler, Jörg/Wang, Huiwen (Eds.): Handbook of Partial Least Squares: Concepts, methods and applications in marketing and related fields. Berlin (2009), (forthcoming).

Ferguson, George A. (1971): Statistical analysis in psychology and education. 3rd ed. New York, NY 1971.

Festge, Fabian/Schwaiger, Manfred (2007): The drivers of customer satisfaction with industrial goods: an international study. In: Advances in International Marketing, Vol. 18 (2007), pp. 179–207.

Fornell, Claes/Larcker, David F. (1981): Evaluating structural equation models with unobservable values and measurement error. In: Journal of Marketing Research, Vol. 18 (1981), No. 1, pp. 39–50.

Fornell, Claes/Cha, Jaesung (1994): Partial Least Squares. In: Bagozzi, Richard P. (Ed.): Advanced methods of marketing research. Cambridge, MA 1994, pp. 52–78.

Fornell, Claes/Johnson, Michael D./Anderson, Eugene W./Cha, Jaesung/Bryant, Barbara E. (1996): The American customer satisfaction index: Nature, purpose, findings. In: Journal of Marketing, Vol. 60 (1996), pp. 7–18.

Fuchs, Sebastian/Sarstedt, Marko (2008): On the use of student samples in major marketing and management research journals. a meta-study. Paper presented at the 32nd Annual Conference of the German Classification Society (GfKl) – Advances in Data Analysis, Data Handling and Business Intelligence, July 16–18, Hamburg.

Gardner, Donald G./Cummings, L. L./Dunham, Randall B./Pierce, Jon L. (1998): Single-item versus multiple-item measurement scales: an empirical comparison. In: Educational and Psychological Measurement, Vol. 58 (1998), No. 6, pp. 898–915.

Geisser, Seymour (1974): A predictive approach to the random effect model. In: Biometrika, Vol. 61 (1974), No. 1, pp. 101–107.

Götz, Oliver/Liehr-Gobbers, Kerstin (2004): Analyse von Strukturgleichungsmodellen mit Hilfe der Partial-Least-Squares(PLS)-Methdode. In: Die Betriebswirtschaft, Vol. 64 (2004), No. 6, pp. 714–738.

Gray, Peter H./Meister, Darren B. (2004): Knowledge sourcing effectiveness. In: Management Science, Vol. 50 (2004), No. 6, pp. 821–834.

Grønholdt, Lars/Martensen, Anne/Kristensen, Kai (2000): The relationship between customer satisfaction and loyalty: Cross-industry differences. In: Total Quality Management, Vol. 11 (2000), No. 4–6, pp. 509–514.

Hartung, Joachim (1989): Statistik. Lehr- und Handbuch der angewandten Statistik. 7th ed. Munich 1989.

Henseler, Jörg/Fassott, Georg (2009): Testing moderating effects with PLS path modeling. An overview over the available approaches. In: Esposito Vinzi, Vincenzo/Chin, Wynne W./Henseler, Jörg/Wang, Huiwen (Eds.): Handbook of Partial Least Squares: Concepts, methods and applications in marketing and related fields. Berlin (2009), (forthcoming).

Henseler, Jörg/Ringle, Christian M./Sinkovics, Rudolf R. (2009): The use of partial least squares path modeling in international marketing. In: Advances in International Marketing, (forthcoming).

Homburg, Christian/Giering, Annette (1998): Konzeptualisierung und Operationalisierung komplexer Konstrukte – Ein Leitfaden für die Marketingforschung. In: Hildebrandt, Lutz/Homburg, Christian (Eds.): Die Kausalanalyse – Instrument der betriebswirtschaftlichen Forschung. Stuttgart 1998, pp. 111–146.

Hulland, John (1999): Use of Partial Least Squares (PLS) in strategic management research: A review of four recent studies. In: Strategic Management Journal, Vol. 20 (1999), No. 2, pp. 195–204.

Jacoby, Jacob (1978): Consumer research: How valid and useful are all our customer behavior research findings? – A state of the art review. In: Journal of Marketing, Vol. 42 (1978), No. 2, pp. 87–96.

Jagodzinski, Wolfgang/Manabe, Kazufumi (2005): Warum Mehrfachindikatoren manchmal auch nicht helfen: Überlegungen zu einem multiplen Indikatormodell für interpersonales Vertrauen im Anschluss an die Anmerkungen von Jürgen Rost. In: ZA-Information, No. 56 (2005), pp. 8–17.

Jöreskog, Karl G. (1970): A general method for analysis of covariance structures. In: Biometrika, Vol. 57 (1970), No. 2, pp. 239–251.

Kwon, Hyungil/Trail, Galen (2005): The feasibility of single-item measures in sport loyalty research. In: Sport Management Review, Vol. 8 (2005), pp. 69–89.

Loo, Robert (2002): A caveat on using single-item versus multiple-item scales. In: Journal of Managerial Psychology, Vol. 17 (2002), No. 1, pp. 68–75.

Nagy, Mark S. (2002): Using a single-item approach to measure facet job satisfaction. In: Journal of Occupational and Organizational Psychology, Vol. 75 (2002), No. 1, pp. 77–86.

Nunnally, Jum C. (1978): Psychometric theory. 2nd ed. New York, NY 1978.

Peter, Jean-Paul (1979): Reliability: A review of psychometric basics and recent marketing practices. In: Journal of Marketing Research, Vol. 16 (1979), No. 1, pp. 6–17.

Rammstedt, Beatrice/Koch, Karina/Borg, Ingwer/Reitz, Tanja (2004): Entwicklung und Validierung einer Kurzskala für die Messung der Big-Five-Persönlichkeitsdimensionen in Umfragen. In: ZUMA-Nachrichten, No. 55 (2004), pp. 5–28. http://www.gesis.org/fileadmin/upload/forschung/publikationen/zeitschriften/zuma_nachrichten/zn_55.pdf, Accessed: December 09, 2008.

Reichheld, Frederick F. (2003): The one number you need to grow. In: Harvard Business Review, Vol. 81 (2003), No. 12, pp. 46–54.

Ringle, Christian M./Wende, Sven/Will, Alexander (2005): SmartPLS 2.0 (M3) beta. http://www.smartpls.de, Accessed: December 09, 2008.

Rossiter, John R. (2002): The C-OAR-SE procedure for scale development in marketing. In: International Journal of Research in Marketing, Vol. 19 (2002), No. 4, pp. 305–335.

Ruekert, Robert W./Curchill, Gilbert A. Jr. (1984): Reliability and validity of alternative measures of channel member satisfaction. In: Journal of Marketing Research, Vol. 21 (1984), No. 2, pp. 226–233.

Ryan, Michael J./Buzas, Thomas/Ramaswamy, Venkatram (1995): Making CSM a power tool – composite indices boost the value of satisfaction measures for decision making. In: Marketing Research, Vol. 7 (1995), No. 3, pp. 11–16.

Sloan, Jeff A./Aaronson, Neil/Cappelleri, Joseph C./Fairclough, Diane L./Varricchio, Claudette/Clinical Significance Consensus Meeting Group (2002): Assessing the clinical significance of single item relative to summated scores. In: Mayo Clinic Proceedings, Vol. 77 (2002), pp. 479–487.

Stone, Mervyn (1974): Cross-validatory choice and assessment of statistical predictions. In: Journal of the Royal Statistical Society, Series B, Vol. 36 (1974), No. 2, pp. 111–148.

Venkatesh, Viswanath/Agarwal, Ritu (2006): Turning visitors into customers: a usability-centric perspective on purchase behavior in electronic channels. In: Management Science, Vol. 52 (2006), No. 3, pp. 367–382.

Wanous, John P./Reichers, Arnon E. (1996): Estimating the reliability of a single-item measure. In: Psychological Reports, Vol. 78 (1996), No. 2, pp. 631–634.

Wanous, John P./Hudy, Michael J. (2001): Single-item reliability: A replication and extension. In: Organizational Research Methods, Vol. 4 (2001), No. 4, pp. 361–375.

Wanous, John P./Reichers, Arnon E./Hudy, Michael J. (1997): Overall job satisfaction: How good are single-item measures? In: Journal of Applied Psychology, Vol. 82 (1997), No. 2, pp. 247–252.

Wold, Herman (1974): Causal flows with latent variables: Partings of the ways in light of NIPALS modeling. In: European Economic Review, Vol. 5 (1974), No. 1, pp. 67–86.

Zeithaml, Valerie A./Berry, Leonard/Parasuraman A. (1996): The behavioral consequences of service quality. In: Journal of Marketing, Vol. 60 (1996), No. 2, pp. 31–46.