HANS BAUMGARTNER and JAN-BENEDICT E.M. STEENKAMP*

In an effort to advance measurement analysis in marketing research, the authors propose three extensions of the current scale development paradigm. First, the authors focus attention on the trait–state distinction and present a model that separates stable sources of substantive variation in a construct from transient sources. Second, the authors develop a classification of measurement error that distinguishes six types of error on the basis of (1) whether the error is stable or transient and (2) whether the error affects individual items, subsets of items, or all items in a scale. They also show how these errors can be modeled using a factor-analytic specification. Third, the authors argue that marketing researchers should make the means of scale items an explicit component of measurement analysis and should test for the invariance of item loadings and intercepts as a prerequisite for meaningful comparisons of scale means across samples and over time. To illustrate the benefits of the proposed procedure, the authors present an extended application of the model to the constructs of brand loyalty and deal proneness.

An Extended Paradigm for Measurement Analysis of Marketing Constructs Applicable to Panel Data

The use of structural equation modeling (SEM) has greatly increased marketing researchers’ concern with the quality of construct measurement and has made sophisticated investigations of validity and reliability standard practice. The starting point for the usual SEM-based measurement analysis is a model in which the observed variables (indicators, items), all measured at a single point in time, are linked to hypothesized underlying constructs (latent variables, factors) as follows (see Bollen 1989):

\[ y_{ij} = \lambda_{ij} \eta_j + \nu_{ij}, \]

where \( y_{ij} \) is a respondent’s observed score on the \( i \)th measure of construct \( j \), \( \eta_j \) is a respondent’s unobserved score on the \( j \)th construct, \( \lambda_{ij} \) is a factor loading relating \( y_{ij} \) to \( \eta_j \), and \( \nu_{ij} \) is an unobserved unique factor score. In addition to the usual statistical assumptions (i.e., \( E[\nu_{ij}] = 0 \) and \( \text{Cov}[\eta_j, \nu_{ij}] = 0 \)), in general, it is assumed that both \( y_{ij} \) and \( \eta_j \) are in deviation form. When multiple measures of each construct are available, it is possible to conduct a comprehensive measurement analysis by assessing unidimensionality, testing convergent and discriminant validity, and computing various indexes of reliability (see, e.g., Gerbing and Anderson 1988; Steenkamp and Van Trijp 1991).

Although the measurement model in Equation 1 has many appealing features, three shortcomings of the specification should be noted. First, temporary and stable aspects of a construct are not distinguished in Equation 1. Second, measure specificity and other sources of systematic but non-construct-related variation in \( y_{ij} \) are confounded with random measurement error. Third, the means of observed and latent variables are not considered explicitly, because all variables are defined as deviations from their means.

Drawing on research on longitudinal factor analysis (e.g., Marsh and Grayson 1994), latent state–trait models (e.g., Steyer, Schmitt, and Eid 1999), and latent curve analysis (e.g., Tisak and Tisak 2000), as well as recent work on different sources of measurement error (e.g., Schmidt, Le, and Ilies 2003), our goal is to develop an integrative procedure...
for conducting measurement analysis that (1) distinguishes temporary and stable components of variance in constructs corresponding to their state and trait aspects, (2) separates systematic sources of nonconstruct variance from random measurement error, and (3) takes into account the means of the observed variables. The approach is an extension of the extant paradigm for scale development in marketing (e.g., Churchill 1979) and requires panel data in which the same observed variables are measured at multiple points in time. This increases the data collection burden, but we show that the benefits are substantial and thus worth the additional effort. Although the procedure may be too demanding for situations in which theory testing is the primary objective, it should be useful, indeed vital, for scale development studies in which a proposed scale is to be subjected to a rigorous construct validation process so that future researchers can use the scale with confidence.

In what follows, we first discuss the three issues in greater detail and show how the conventional measurement model in Equation 1 can be extended to incorporate these features. We then present an extended application of the proposed procedure to the substantive area of brand loyalty and deal proneness. The article ends with a discussion of the findings and recommendations for further research.

THE NEED TO EXTEND THE CONVENTIONAL MEASUREMENT MODEL

Stable and Transient Components of a Construct

Researchers in the behavioral sciences have found it useful to distinguish between relatively stable dispositions of people (traits) and transitory characteristics of individuals (states) that change with circumstances and over time (for a discussion, see Chaplin, John, and Goldberg 1988). Perhaps the best example of the trait–state distinction in marketing can be found in research on involvement, in which many authors have distinguished between enduring and situational sources of personal relevance (e.g., Celsi and Olson 1988). As another example, in work on optimum stimulation level, it is important to know whether a person’s current level of stimulation (state stimulation) deviates from his or her optimum stimulation level (trait stimulation), because people are predicted to engage in exploratory behavior when a deviation occurs (e.g., Steenkamp and Baumgartner 1992). Two other examples are the differences between dispositional and situational self-regulatory focus and life satisfaction.

We believe that the distinction between states and traits is important in many areas of marketing research and that the scale development process should include an assessment of the extent to which the instrument captures trait or state aspects. Unfortunately, the trait–state distinction has been neglected in marketing research, and when scale developers have investigated the stability of an instrument, they have often used problematic methods, such as test–retest correlations. For example, when we reviewed the 192 scale development efforts that Bearden and Netemeyer (1999) describe, we encountered both cases in which a construct was viewed as a trait and no evidence of stability was provided and instances in which a construct was defined as a state and lack of stability was never assessed. Furthermore, although Bearden and Netemeyer report a test–retest corre-

lation for 23% of the scales, it is often unclear whether it was intended as a measure of stability and/or test–retest reliability.

Systematic and Random Sources of Measurement Error

One of the major advantages of using SEM is that the total variance in an observed measure can be decomposed into various sources of variation and that not all the variance is treated as substantive (construct-relevant, content-based) variation. Traditionally, the emphasis has been on partitioning the observed variance into just two components: substantive variance and (random) error variance. This practice is problematic for several reasons.

First, in general, treating all sources of nonsubstantive variation as an undifferentiated amalgam of measurement error gives the misleading impression that the causes of measurement error are unknown and unknowable (Schmidt and Hunter 1999). Second, unless different forms of measurement error are modeled explicitly, it is impossible to obtain insights into the relative importance of the different sources of error and take ameliorative action. Third, observed variables can covary for reasons other than those caused by substantive factors (Podsakoff et al. 2003). If these systematic error correlations are ignored, misleading conclusions may be derived from the data, as we show subsequently.1

To facilitate a more explicit treatment of sources of error, we propose a classification of measurement error along two dimensions. The first dimension distinguishes among three types of error, depending on how many items are affected: (1) measurement error that is specific to a single item within a scale, (2) measurement error that affects a subset of items within a scale, and (3) measurement error that affects all the items in a scale. The second dimension distinguishes between sources of error that are restricted to a given occasion of measurement (transient) and those that are consistent over time (stable).

Item-specific error. Some forms of measurement error are specific to individual items within a scale, in the sense that the error does not induce dependencies among different items. Usually, this is the only type of error that is accounted for in SEM analyses.

“Transient item-specific error” includes all sources of measurement error that are restricted to one occasion of measurement and, for practical purposes, can be treated as random measurement error. Examples include temporary personal (e.g., lapses of attention) or situational (e.g., a distracting noise) influences that affect only a single item. In contrast, “stable item-specific error” persists over time. The distinction is similar to one made in the factor-analytic literature, in which unique variance (as opposed to common variance) is sometimes partitioned into measure specificity and random measurement error. Usually, the two sources of variance are confounded, but when repeated measurements

1One of the few areas in which systematic nonsubstantive sources of covariation among observed measures have been studied explicitly is multitrait–multimethod analysis, in which they are referred to as “method effects” (e.g., Cote and Buckley 1987). However, this literature is often focused more on the statistical issues arising in these types of analyses than on learning something substantively interesting about the construct of interest or measurement error (Cronbach 1995).
are available for the same respondents, separate variance components can be estimated. Although stable item-specific error induces dependencies between identical items measured at different points in time, it does not lead to error correlations between different items, and therefore it acts like random error by attenuating relationships between constructs. Likely sources of this error are people’s idiosyncratic reactions to the way an item is worded (e.g., a certain term may have special meaning for some respondents) and the unique response scale associated with a particular item.

**Item-subset error.** Some forms of error affect a subset of items within a scale and therefore lead to correlated errors of measurement among the items involved, often resulting in an inflation of scale reliability. If the error also affects similar items in other scales, as is likely to be the case, correlations between constructs can be seriously biased.

As in the case of item-specific error, a common source of shared nonsubstantive variation among a subset of items is wording effects, though with item-subset error, similar terminology or a particular response scale must be used in several different items. One of the most important types of wording effects is positive or negative item wording (Wong, Rindfleisch, and Burroughs 2003). Another instance of item-subset error may arise when the items indicating a construct come from different raters or informants (Bollen and Paxton 1998).

Item-subset errors can be transient or stable. For example, assume that a scale consists of five positively and five negatively worded items and that acquiescence leads to error correlations among the items worded in the same direction (Baumgartner and Steenkamp 2001). If acquiescence is viewed as a personality trait, the error should be stable over time, resulting in “stable item-subset error.” Conversely, if acquiescence is attributed to characteristics of the situation (e.g., time pressure), it may be restricted to a particular occasion of measurement, which implies that there is “transient item-subset error.”

**Scalewide error.** Scalewide error affects all the items in a scale, and it can seriously distort the findings of measurement analyses (e.g., it can exaggerate the amount of substantive variance in an item and inflate the reliability of the scale). If the sources of scalewide error also affect other scales, estimated relationships with other constructs will be biased as well. The damaging effects of shared method variance have been acknowledged in the industrial and organizational psychology literature, in which it has been a cause of concern that all ratings are often provided by the same source (e.g., Podsakoff et al. 2003).

Schmidt and Hunter (1999) and Schmidt, Le, and Illies (2003) discuss the notion of “transient scalewide error” from a reliability perspective, and Schmidt and Hunter (1999, p. 193) suggest that this error is caused by “variations in mood, feeling, mental efficiency, or general mental state.” For example, research shows that a person’s mood can bias judgments in a mood-congruent direction (Gardner 1985). This implies that differences in respondents’ mood states can be a temporary source of shared nonsubstantive variation in ratings at the scale level.

Errors affecting all items in a scale can also be stable, which leads to “stable scale-wide error.” Many scales used in marketing do not contain reverse-scored items, which implies that response styles such as acquiescence may induce scalewide error (Baumgartner and Steenkamp 2001). Other dispositional variables often implicated in method effects include social desirability (Mick 1996) and negative affectivity (Podsakoff et al. 2003).

**Comparison of Construct Means Across Groups and Over Time**

Scales are often used to ascertain whether groups of entities differ in their standing on a construct or whether entities have changed over time. It is common to compare construct means across groups, and sometimes, the construct validation process itself includes a comparison of scale means across groups that are known or assumed to differ on the construct of interest. For example, Tian, Bearden, and Hunter (2001) show that tattoo and body-piercing artists, owners of customized low-rider automobiles, and members of a medievist reenactment group had significantly higher scores on their need-for-uniqueness scale than members of a heterogeneous mail survey sample. In longitudinal research, trajectories of constructs are frequently studied. For example, a key question in work based on the American Customer Satisfaction Index (Fornell et al. 1996) is how satisfaction has developed over time and what factors account for changes in satisfaction.

Although comparisons of means are reported frequently in published studies, many researchers are apparently not aware that these comparisons are meaningful only if certain measurement assumptions are met. Several authors have discussed the necessary conditions of measurement equivalence (e.g., Steenkamp and Baumgartner 1998), but it is rare for authors to conduct the required invariance tests before they embark on the substantive comparisons of interest. Instead of trying to get researchers to test for measurement invariance every time they use a certain scale, it is preferable to ask scale developers to examine whether the proposed scale is suitable for conducting comparisons of means across groups and over time. It is clear that the observed and latent variables cannot be specified as deviations from their means when the means are of explicit interest. Thus, the conventional measurement model, which ignores information about the means, must be extended to incorporate the variable means.

**A MULTI-ITEM, MULTICONSTRUCT, MULTI-OCCASION MEASUREMENT MODEL**

When respondents complete the same set of items assessing several different constructs on at least three occasions, the conventional measurement model can be expanded to include stable and transient sources of substantive and error variation as well as the means of the variables. We specify the model as follows:

\[
y_{ijt} = \tau_{ijt} + \lambda_{ijt} \eta_{jt} + \omega_{ijt},
\]

where \(y_{ijt}\) is a person’s observed score on the \(i\)th measure of construct \(j\) at time \(t\), \(\eta_{jt}\) is an unobserved occasion-specific measure of construct \(j\) at time \(t\), \(\lambda_{ijt}\) is a coefficient (factor loading) relating \(y_{ijt}\) to \(\eta_{jt}\) (for identification, one loading per factor is set to one), \(\omega_{ijt}\) is a composite of all the errors affecting \(y_{ijt}\), and \(\tau_{ijt}\) is an intercept term. As usual, we assume that \(\eta_{jt}\) is uncorrelated with \(\omega_{ijt}\) so that the total variation in \(y_{ijt}\) consists of two orthogonal components: sub-
stantive (or construct-relevant) variation (captured by $\eta_{jt}$) and error (nonsubstantive, construct-irrelevant) variation (the terms contained in $\sigma_{ijt}$). Subsequently, we describe (1) how $\eta_{jt}$ is specified to incorporate the distinction between traits and states, (2) how $\sigma_{ijt}$ is decomposed to include the various sources of error, and (3) how the means of the variables are modeled.

**Modeling States and Traits**

Steyer, Schmitt, and Eid (1999) demonstrate how traits and states can be distinguished using a second-order factor model, and we adopt their procedure. Thus,

$$
\eta_{jt} = \alpha_{jt} + \beta_{jt} \eta_{jt} + \xi_{jt},
$$

where $\eta_{jt}$ is a person’s score on the stable or trait component of the $j$th construct ($S$ stands for stable), $\eta_{jt} = \beta_{jt}$ is the corresponding transient or state component ($T$ stands for transient), $\beta_{jt}$ is a second-order factor loading relating $\eta_{jt}$ to $\eta_{jt}$, and $\alpha_{jt}$ is an intercept term. Equation 3 shows that the occasion-specific factor $\eta_{jt}$ consists of a trait component that is time invariant and a state component that leads to deviations from the trait score at time $t$. We assume that $\eta_{jt}$ is uncorrelated with $S$ and that $\xi_{jt}$, associated with different $t$, are also uncorrelated, though the state components of different constructs $j$ can be correlated for a given $t$ and different trait components can be freely correlated.

Although the trait components are uncorrelated with the state components, the occasion-specific effects $\eta_{jt}$ are correlated, and their stability can be ascertained. Specifically, it can be shown that the squared correlation between $\eta_{jt}$ and $\eta_{jt}$ is equal to the proportion of trait variation in $\eta_{jt}$ times the proportion of trait variation in $\eta_{jt}$. Therefore, when all the substantive variation is trait variation, the correlation between $\eta_{jt}$ and $\eta_{jt}$ will be perfect; when none of the substantive variation is trait variation, it will be zero. The proportion of trait variation in $\eta_{jt}$ is given by

$$
\rho_{jt} = \frac{(\beta_{jt})^2 \text{Var}(\eta_{jt})}{(\beta_{jt})^2 \text{Var}(\eta_{jt}) + \text{Var}(\xi_{jt})},
$$

and the proportion of state (residual) variation is equal to 1 minus this expression. For a pure trait measure, all the variation in $\eta_{jt}$ should be trait variation; for a pure state measure, all the variation should be state variation. In practice, it is unlikely that a scale will be a pure trait or pure state measure, but, at a minimum, the proportion of trait variation should be greater than .5 for a trait measure and less than .5 for a state measure.

**Modeling Random and Systematic Sources of Measurement Error**

A major benefit of the proposed model is that the various sources of measurement error, both random and systematic, are modeled explicitly. Specifically, we propose the following specification of the error structure for the composite error component in Equation 2:

$$
\sigma_{ijt} = \xi_{SSWE} + \xi_{TSWE} + I_{ijt} \in S + \sigma_{ijt}.
$$

and the proportion of state (residual) variation is equal to 1 minus this expression. For a pure trait measure, all the variation in $\eta_{jt}$ should be trait variation; for a pure state measure, all the variation should be state variation. In practice, it is unlikely that a scale will be a pure trait or pure state measure, but, at a minimum, the proportion of trait variation should be greater than .5 for a trait measure and less than .5 for a state measure.

error, $\xi_{SSWE}$ refers to transient (item)-subset error, $\xi_{SIE}$ indicates stable item-(specific) error (there will be as many error terms as there are distinct items), $\epsilon_{ijt}$ is transient item-specific (i.e., random) error, $\lambda_{SIE}$ is a factor loading, and $I$ is an indicator variable that takes the value of 1 if the ith item of construct $j$ at time $t$ belongs to subset $S$ and 0 if otherwise. Although we show only one type of (stable or transient) item-subset error in Equation 5, multiple subset factors can be specified if necessary (e.g., separate error factors for positively and negatively worded items may be included).

We assume that all errors are uncorrelated with the substantive effects, have means of zero, and are mutually uncorrelated.2 For identification, one loading $\lambda_{SIE}$ per factor must be set to one. Note that we have assumed that all items $y_{ijt}$ load equally on $\xi_{SSWE}$, $\xi_{TSWE}$, $\xi_{SIE}$, and $\xi_{IIE}$. If these restrictions are inappropriate (based on significant modification indexes and substantial expected parameter changes), they can be relaxed, though a model in which all error loadings are freely estimated is not identified. It is difficult to provide general identification rules for models of this complexity, but in general, a model with a minimum of two constructs measured by at least four indicators at each of three points in time is identified. In practice, identification must often be ascertained empirically (see Bollen 1989).

Equation 5, which models all sources of measurement error as additive first-order factors, is our recommended specification for the error structure. Two other possibilities have been suggested in the literature. One is the correlated uniqueness (CU) model (Marsh 1989), which allows for a multidimensional error structure. However, because a lack of convergence does not appear to be a serious problem with the proposed specification, at least at larger sample sizes (better convergence is the primary benefit of the CU model, but it is less of a problem in our case because the error factors are specified to be uncorrelated), there are few reasons to recommend the CU model (for a discussion, see Lance, Nobel, and Scullen 2002). The other alternative is the direct product model (Browne 1984), which allows for interactions between traits and methods. However, as Podskoff and colleagues (2003) discuss, the direct product model has some serious shortcomings (e.g., it cannot be used to evaluate the measurement quality of individual items), which argues against its use in modeling measurement error.

**Variance Decompositions and Reliability**

The full measurement model for $y_{ijt}$ can be obtained by combining Equations 2, 3, and 5. Under the assumptions we made previously, this yields the following expression for the total variance of $y_{ijt}$:

$$
\sigma_{ijt}^2 = \lambda_{ijt}^2 \beta_{ijt}^2 \text{Var}(\eta_{ijt}) + \lambda_{ijt}^2 \text{Var}(\xi_{ijt}) + \text{Var}(\xi_{SSWE}) + \text{Var}(\xi_{TSWE}) + \text{Var}(\xi_{SIE}) + \text{Var}(\xi_{IIE}).
$$

2The assumptions that all errors have zero means (nonzero means can be absorbed into the intercepts) and are uncorrelated with the substantive effects are standard assumptions in factor-analytic modeling (e.g., Bollen 1989). If the assumption about the uncorrelatedness of the various error factors is violated, it will be reflected in bad model fit.
The first two terms on the right-hand side of Equation 6 contain the two substantive sources of variance, and the last six terms contain the six error components. By expressing each component of variation as a proportion of the total variation, we can partition the total variance in \( y_{ijt} \) into substantive variance due to traits and states and into error variance due to stable and transient scalewide error, stable and transient item-subset error, stable item-specific error, and random error. Researchers can use these variance proportions to diagnose the quality of individual items and eliminate poorly performing items. To report the results succinctly, the individual variance proportions can also be averaged by construct and time period; in the literature, these averages are referred to as “average variance extracted” (Fornell and Larcker 1981).

It is also possible to compute indexes of reliability analogous to coefficient alpha. In general, reliability is defined as the proportion of observed score variance accounted for by true scores. Depending on the researcher’s objectives, different variance components may be counted as true scores (e.g., Steyer, Schmitt, and Eid 1999). Here, we discuss three coefficients of composite reliability (assuming a unit-weighted composite of scale items) based on the following assumptions: (1) True score variance should include only substantive (construct-relevant) variance, and (2) depending on the context, a researcher will be interested in trait variance, state variance, or both. The resulting three reliabilities, which we call “trait reliability” (\( \rho_{S,jt} \)), “state reliability” (\( \rho_{T,jt} \)), and “substantive reliability” (\( \rho_{Sub,jt} \)), are given as follows:

\[
\rho_{S,jt} = \frac{\left( \frac{\sum_{i=1}^{N_j} \lambda_{ij}^2}{s_j} \right)^2 \beta_j^2 \text{Var}(\eta_j^S)}{\sigma_j^2},
\]

\[
\rho_{T,jt} = \frac{\left( \frac{\sum_{i=1}^{N_j} \lambda_{ij}^2}{s_j} \right)^2 \text{Var}(\zeta_{jt})}{\sigma_j^2}, \text{ and}
\]

\[
\rho_{Sub,jt} = \rho_{S,jt} + 2 \rho_{T,jt},
\]

where the denominator in Equations 7 and 8, the total variance of an unweighted composite of all items measuring construct \( j \), is given by

\[
\sigma_j^2 = \left( \frac{\sum_{i=1}^{N_j} \lambda_{ij}^2}{s_j} \right)^2 \beta_j^2 \text{Var}(\eta_j^S) + \left( \frac{\sum_{i=1}^{N_j} \lambda_{ij}^2}{s_j} \right)^2 \text{Var}(\zeta_{jt}) + \sum_{i=1}^{N_j} \lambda_{ij}^2 \text{Var}(\epsilon_{ijt}),
\]

where \( N_j \) is the total number of items measuring construct \( j \) and \( s_j \) is the number of items measuring construct \( j \) for which an item-subset error is specified (\( s_j < N_j \)).

Modeling the Means of the Constructs

The model for the construct means can be obtained from Equation 3. If the means of the trait factors \( \eta_j^S \), \( E(\eta_j^S) \), are denoted by \( \kappa_j^S \), the means of the occasion-specific factors \( \eta_{jt} \), \( E(\eta_{jt}) \), are given by

\[
E(\eta_{jt}) = \alpha_{jt} + \beta_j \kappa_j^S.
\]

To identify the means part of the model, the minimum requirements are that (1) the measurement intercepts \( \tau_{ijt} \) associated with the marker items for the occasion-specific factors \( \eta_{jt} \) (i.e., those items whose loadings \( \lambda_{ijt} \) on \( \eta_{jt} \) are set to one) are zero and (2) the equation intercepts \( \alpha_{jt} \) associated with the marker factors for the trait factors \( \eta_j^S \) (i.e., the occasion-specific factors whose loadings \( \beta_j \) on \( \eta_{jt} \) are set to one) are also zero. Under this specification, the mean of the occasion-specific factor that serves as the marker item is given by \( \kappa_j^S \) (this is usually the mean of the occasion-specific factor for the first period; i.e., \( E(\eta_{j1}) = \kappa_j^S \)), and the means of the other occasion-specific factors are given by \( \alpha_{jt} + \beta_j \kappa_j^S \).

To test for the equality of the means of the occasion-specific factors, the restrictions that \( \alpha_{jt} + \beta_j \kappa_j^S = \kappa_j^S \) are imposed. If the chi-square value for the constrained model is significantly larger than the one for the unconstrained model, the hypothesis of stability is rejected, and the reason for the lack of stability can be investigated using more specific comparisons of means.

If data for multiple groups of respondents are available, it is possible to compare means across samples. In this case, it is possible to test whether the means of the occasion-specific factors are the same for a given \( t \) (i.e., whether \( E(\eta_{jt}^{G1}) = E(\eta_{jt}^{G2}) \), where \( G_1 \) and \( G_2 \) denote two different groups of respondents) or whether the means of the trait factors are invariant across groups (i.e., whether \( \kappa_j^{G1} = \kappa_j^{G2} \)).

Regardless of whether construct means are compared over time or across samples, certain invariance constraints must be imposed on the model before valid comparisons of means can be conducted. The minimal restrictions for identifying the means part of the model (as discussed previously) are not sufficient to ensure comparability of measurements. The issues are essentially the same as those discussed in the literature on multisample analysis (e.g., Steenkamp and Baumgartner 1998). Specifically, meaningful comparison of construct means requires both metric and scalar invariance. Metric invariance means that the first-order factor loadings \( (\lambda_{ijt}, \lambda_{ij}^{SIE}) \) are the same. Scalar invariance means that the item intercepts \( (\tau_{ijt}) \) are also the same. For comparisons over time, the loadings and intercepts for a given item should be invariant over time (i.e., \( \lambda_{ijt} = \lambda_{ij}^{SIE} = \lambda_{ij}^{SIE.G1} \), and \( \tau_{ijt} = \tau_{ij} \)); for comparisons across samples, the loadings and intercepts for a given time should be invariant across samples (i.e., \( \lambda_{ijt}^{G1} = \lambda_{ijt}^{G2} = \lambda_{ij}^{SIE.G1} = \lambda_{ij}^{SIE.G2} \), and \( \tau_{ijt}^{G1} = \tau_{ijt}^{G2} \)). If across-group over-time comparisons are to be conducted simultaneously, both types of invariance must be satisfied. As Steenkamp and Baumgartner (1998) argue, meaningful comparisons of means are possible as long as at least two items per factor have invariant loadings and intercepts over time or across samples.

AN ILLUSTRATIVE APPLICATION OF THE PROPOSED MODEL TO THE CONSTRUCTS OF BRAND LOYALTY AND DEAL PRONENESS

In this section, we apply the proposed model to scales that measure the traits of brand loyalty and deal proneness.
We show how the procedure can be used to assess the measurement quality of individual items and the scale as a whole, and we also test the measurement invariance of the scales over time and compare the means of the occasion-specific brand loyalty and deal proneness factors across three periods.

Method

The scale for measuring the brand loyalty trait is based on the work of Beatty and Kahle (1988) and Raju (1980) and consists of five items, two of which are negatively worded (Items 2 and 4). The scale for measuring the trait of deal proneness is based on the work of Lichtenstein, Netemeyer, and Burton (1995) and consists of four items, one of which (Item 2) is negatively worded. The global market research company GfK routinely administers these nine items through mail questionnaires to its Dutch household panel as part of an annual update on panel attitudes and dispositions. We obtained these data for three separate occasions (in early 2000, 2002, and 2003). The effective sample size over the period of three years was 1991 consumers. We assessed the items using five-point Likert scales (1 = “completely disagree,” and 5 = “completely agree”).

To represent the substantive variance in the nine items, we specified six occasion-specific factors, with trait variance modeled as a second-order factor and state variance treated as residual variation. We included all six sources of error and specified separate item-subset factors for the positively and negatively worded items.

Preliminary analyses indicated that the data were well behaved. The maximum absolute values of univariate skewness and kurtosis were .76 and .91, respectively, and the coefficient of relative multivariate kurtosis was 1.13, well below levels that would raise concerns about the distributional assumptions (Kline 2005). Therefore, we chose normal-theory maximum likelihood as the method of estimation. We performed all analyses in LISREL 8.7, using a variance–covariance matrix and variable means as input.

Results

Model specification. We recommend the following sequential model comparison procedure when conducting measurement analyses. Initially, the baseline model of Equation 1 is estimated, in which first-order correlated occasion-specific factors represent the constructs of interest and the error (unique) factors are specified to be uncorrelated. In general, this model fits poorly. Stable item-specific error factors are added next (one for each item) because they tend to improve the fit of the model substantially. Then, stable item-subset, transient item-subset, stable scalewide, and transient scalewide error factors are included sequentially, provided that they each account for a significant portion of the variance in the ratings. After the error specification is in place, the distinction between traits and states can be added (this second-order factor specification is usually more restrictive than the first-order model). Finally, the invariance of the loadings and item intercepts is assessed, and if measurement invariance is satisfied, the stability of the construct means can be investigated.

Table 1 reports these model comparisons for the current context. In addition to the chi-square statistic, we report three stand-alone indexes (root mean square error of approximation [RMSEA], standardized root mean square residual [SRMR], and consistent Akaike information criterion [CAIC]) and two incremental fit indexes (comparative fit index [CFI] and Tucker–Lewis index [TLI]), which have proved useful for evaluating overall model fit (e.g., Steenkamp and Baumgartner 1998). To compare models, we rely on the fit indexes that trade off goodness of fit and model complexity (RMSEA, CAIC, and TLI), especially CAIC, because it has been shown to be the most effective index for distinguishing between correctly and incorrectly specified models (Williams and Holahan 1994).

The baseline model (M1) fit the data poorly, and the introduction of stable item-specific factors (M2) improved the fit dramatically. Many researchers would probably be satisfied with the overall fit of this model (all fit indexes are better than commonly used rule-of-thumb criteria) and not introduce additional error factors. However, as we show subsequently, this would be a serious mistake.

Next, we added two stable item-subset factors (M3) and six transient item-subset factors (M4) to model shared variance between the three negatively worded items (two items for brand loyalty and one for deal proneness) and between the six positively worded items (three items each for brand loyalty and deal proneness). Because the variances of the transient error factors for the negatively worded items were very small and nonsignificant, we eliminated these factors (M5). This simplification did not decrease the fit of the model in terms of RMSEA and TLI and actually improved the fit in terms of CAIC. Then, we included one stable (M6) and three transient (M7) scalewide error factors, but the transient factors were unnecessary because the variances were small and nonsignificant, RMSEA and TLI did not improve, and CAIC actually increased (thus, M6 is the preferred first-order specification).

Finally, we modeled the distinction between the trait and the state aspects of brand loyalty and deal proneness using the second-order specification of Equation 2. This model (M8) fit the data as well as any of the previous models based on RMSEA and TLI and had the lowest CAIC. Furthermore, the fit was also excellent in an absolute sense: $\chi^2(281) = 676.50$, RMSEA = .027, SRMR = .035, CFI = .982, and TLI = .978. The LISREL code for this model appears in the Appendix.

Diagnosis of measurement quality. Table 2 shows the variance decompositions for each of the brand loyalty and deal proneness items into the two substantive sources of variation (trait and state) and the six error sources. Table 2 also reports averages across items and periods. On average, 39% of the variance in the brand loyalty items and 28% of the variance in the deal proneness items are substantive variance. These values are below the 50% level that Fornell and Larcker (1981) recommend. Most of the substantive

---

3The brand loyalty scale consists of the following five items: “After I get used to a brand, I hate to switch,” “When another brand is on sale, I generally purchase it rather than my usual brand,” “I feel really committed to the brands I buy,” “If my preferred brand were not available at the store, it would make little difference to me if I had to choose another brand,” and “Even though certain products are available in a number of different brands, I always tend to buy the same brand.” The deal proneness scale consists of the following four items: “I enjoy buying products that are on offer,” “In general, I do not respond to promotional offers,” “Buying brands on offer makes me happy,” and “I love special promotional offers.”

4A SAS IML program, which performs the variance decompositions and reliability analyses on the basis of LISREL output, can be obtained on request.
Table 1

MODEL COMPARISONS FOR THE BRAND LOYALTY AND DEAL PRONENESS SCALES

<table>
<thead>
<tr>
<th>Model</th>
<th>Minimum Fit Function</th>
<th>d.f.</th>
<th>RMSEA</th>
<th>SRMR</th>
<th>CAIC</th>
<th>CFI</th>
<th>TLI</th>
</tr>
</thead>
<tbody>
<tr>
<td>(M1) Freely correlated occasion-specific first-order factors with uncorrelated uniquenesses</td>
<td>3560.54</td>
<td>309</td>
<td>.073</td>
<td>.065</td>
<td>4154</td>
<td>.855</td>
<td>.835</td>
</tr>
<tr>
<td>(M2) M1 plus stable item-specific errors</td>
<td>969.43</td>
<td>282</td>
<td>.035</td>
<td>.052</td>
<td>1795</td>
<td>.969</td>
<td>.962</td>
</tr>
<tr>
<td>(M3) M2 plus stable item-subset errors for positively and negatively worded items</td>
<td>725.37</td>
<td>280</td>
<td>.028</td>
<td>.039</td>
<td>1568</td>
<td>.980</td>
<td>.975</td>
</tr>
<tr>
<td>(M4) M3 plus transient item-subset errors for positively and negatively worded items</td>
<td>668.46</td>
<td>274</td>
<td>.027</td>
<td>.039</td>
<td>1562</td>
<td>.982</td>
<td>.977</td>
</tr>
<tr>
<td>(M5) M4 minus transient item-subset errors for negatively worded items</td>
<td>677.41</td>
<td>277</td>
<td>.027</td>
<td>.039</td>
<td>1546</td>
<td>.982</td>
<td>.977</td>
</tr>
<tr>
<td>(M6) M5 plus stable scalewide error</td>
<td>662.48</td>
<td>276</td>
<td>.027</td>
<td>.034</td>
<td>1539</td>
<td>.983</td>
<td>.978</td>
</tr>
<tr>
<td>(M7) M6 plus transient scalewide errors</td>
<td>657.46</td>
<td>273</td>
<td>.027</td>
<td>.034</td>
<td>1560</td>
<td>.983</td>
<td>.978</td>
</tr>
<tr>
<td>(M8) M6 with second-order factor specification for traits and states</td>
<td>676.50</td>
<td>281</td>
<td>.027</td>
<td>.035</td>
<td>1510</td>
<td>.982</td>
<td>.978</td>
</tr>
<tr>
<td>(M9) M8 with fully invariant factor loadings</td>
<td>732.85</td>
<td>313</td>
<td>.026</td>
<td>.037</td>
<td>1292</td>
<td>.981</td>
<td>.979</td>
</tr>
<tr>
<td>(M10) M9 with fully invariant item intercepts</td>
<td>808.87</td>
<td>327</td>
<td>.027</td>
<td>.037</td>
<td>1247</td>
<td>.978</td>
<td>.977</td>
</tr>
<tr>
<td>(M11) M10 with partially invariant item intercepts</td>
<td>760.09</td>
<td>325</td>
<td>.026</td>
<td>.037</td>
<td>1216</td>
<td>.981</td>
<td>.979</td>
</tr>
<tr>
<td>(M12) M11 with stable construct means over time</td>
<td>782.08</td>
<td>329</td>
<td>.026</td>
<td>.037</td>
<td>1203</td>
<td>.980</td>
<td>.978</td>
</tr>
</tbody>
</table>

Variance is trait variance (85% in the case of brand loyalty and 75% in the case of deal proneness), so these scales are clearly trait scales. This is also confirmed by a variance decomposition of the $\eta_{jt}$ using Equation 4, which shows that, on average, 86% and 77% of the variation in the occasion-specific brand loyalty and deal proneness factors are trait variation. The first brand loyalty item (“Once I get used to a brand, I hate to switch”) and the fourth deal proneness item (“I love special promotional offers”) are the best measures of brand loyalty and deal proneness, respectively, because they contain the highest portion of trait variance. The wording of these items is problematic for Dutch consumers. In contrast, the first brand loyalty item does particularly well because it contains little item-specific variance. The remaining error is primarily due to stable scalewide variance (8% of total error for both brand loyalty and deal proneness) and item-subset variance (again, 8% of total error in both cases). The transient sources of these errors are negligible.

The composite reliabilities of the scales as a whole are as follows: .56, .63, and .61 for the trait reliability of brand loyalty in Periods 1, 2 and 3; .36, .40, and .37 for the trait reliability of deal proneness; .16, .06, and .09 for the state reliability of brand loyalty; .18, .08, and .09 for the state reliability of deal proneness; .72, .68, and .70 for the substantive reliability of brand loyalty; and .54, .48, and .46 for the substantive reliability of deal proneness. Because the scales are trait scales, the low state reliabilities are unprob-
lematic and, indeed, even desirable. However, the trait reliabilities are marginal, particularly for deal proneness. Either better items or longer scales are needed to achieve adequate levels of trait reliability. If the distinction between traits and states is of no concern, attention can focus on substantive reliability, which is adequate in the case of brand loyalty.

It is instructive to compare these findings with the results of a conventional SEM measurement analysis. We use model M2 in Table 1 for this purpose. The substantive conclusions are the same, whether we use model M1 or separate two-factor models for each period. As we already noted, many researchers would probably deem the fit of model M2 to be adequate. More detailed analyses would indicate that Items 1, 3, and 5 for the brand loyalty scale and Items 1 and 4 for the deal proneness scale contain higher portions of substantive variance. These conclusions are consistent with the prior findings. However, the conventional analysis is deficient in the following respects: First, it exaggerates the reliability of the measures. For example, the average item-level substantive variance extracted figures for brand loyalty and deal proneness are .46 and .38 rather than .39 and .28. The estimated composite (analogous to substantive) reliabilities (averaged over time) are .82 and .70 for brand loyalty and deal proneness rather than .70 and .49 (the average coefficient alphas were .81 and .68). This occurs because item-subset errors and scalewide errors mistakenly contribute to substantive covariation. Second, the conventional analysis provides few insights into why some items perform better or worse than others.

The relationship between brand loyalty and deal proneness. The proposed specification makes it possible to clearly distinguish between the trait and the state elements of a construct. The estimate of the trait correlation is –.70, whereas the estimates of the state correlations in Periods 1–3 are –.25, –.11, and –.40. The correlations between the occasion-specific brand loyalty and deal proneness constructs are –.58, –.63, and –.65 for Periods 1–3. Thus, the trait correlation between brand loyalty and deal proneness is negative (as would be expected) and substantial, and the correlations between the occasion-specific factors are somewhat attenuated by the lower state correlations.

Again, it is revealing to compare these results with the findings from a conventional SEM analysis in which traits and states are not distinguished and systematic error is

---

### Table 2

**VARIANCE DECOMPOSITIONS FOR THE BRAND LOYALTY (BL) AND DEAL PRONENESS (DP) SCALES**

<table>
<thead>
<tr>
<th></th>
<th>Trait Variance</th>
<th>State Variance</th>
<th>Substantive Variance</th>
<th>Stable Scalewide Variance</th>
<th>Transient Scalewide Variance</th>
<th>Stable Item-Subset Variance</th>
<th>Transient Item-Subset Variance</th>
<th>Stable Item-Specific Variance</th>
<th>Random Error Variance</th>
<th>Total Error Variance</th>
</tr>
</thead>
<tbody>
<tr>
<td>BL1</td>
<td>.39</td>
<td>.11</td>
<td>.50</td>
<td>.04</td>
<td>.00</td>
<td>.05</td>
<td>.02</td>
<td>.07</td>
<td>.32</td>
<td>.50</td>
</tr>
<tr>
<td>BL2</td>
<td>.29</td>
<td>.08</td>
<td>.37</td>
<td>.04</td>
<td>.00</td>
<td>.02</td>
<td>.00</td>
<td>.09</td>
<td>.49</td>
<td>.63</td>
</tr>
<tr>
<td>BL3</td>
<td>.32</td>
<td>.09</td>
<td>.41</td>
<td>.04</td>
<td>.00</td>
<td>.06</td>
<td>.02</td>
<td>.10</td>
<td>.37</td>
<td>.59</td>
</tr>
<tr>
<td>BL4</td>
<td>.25</td>
<td>.07</td>
<td>.32</td>
<td>.05</td>
<td>.00</td>
<td>.02</td>
<td>.00</td>
<td>.17</td>
<td>.45</td>
<td>.68</td>
</tr>
<tr>
<td>BL5</td>
<td>.31</td>
<td>.09</td>
<td>.40</td>
<td>.05</td>
<td>.00</td>
<td>.06</td>
<td>.02</td>
<td>.08</td>
<td>.39</td>
<td>.60</td>
</tr>
<tr>
<td>BL6</td>
<td>.43</td>
<td>.04</td>
<td>.47</td>
<td>.05</td>
<td>.00</td>
<td>.06</td>
<td>.02</td>
<td>.06</td>
<td>.35</td>
<td>.53</td>
</tr>
<tr>
<td>BL7</td>
<td>.32</td>
<td>.03</td>
<td>.35</td>
<td>.05</td>
<td>.00</td>
<td>.02</td>
<td>.00</td>
<td>.08</td>
<td>.50</td>
<td>.65</td>
</tr>
<tr>
<td>BL8</td>
<td>.37</td>
<td>.03</td>
<td>.40</td>
<td>.05</td>
<td>.00</td>
<td>.06</td>
<td>.02</td>
<td>.16</td>
<td>.31</td>
<td>.60</td>
</tr>
<tr>
<td>BL9</td>
<td>.26</td>
<td>.02</td>
<td>.29</td>
<td>.05</td>
<td>.00</td>
<td>.02</td>
<td>.00</td>
<td>.17</td>
<td>.47</td>
<td>.71</td>
</tr>
<tr>
<td>BL10</td>
<td>.30</td>
<td>.03</td>
<td>.33</td>
<td>.06</td>
<td>.00</td>
<td>.07</td>
<td>.02</td>
<td>.15</td>
<td>.37</td>
<td>.67</td>
</tr>
<tr>
<td>BL11</td>
<td>.45</td>
<td>.07</td>
<td>.52</td>
<td>.05</td>
<td>.00</td>
<td>.06</td>
<td>.02</td>
<td>.06</td>
<td>.29</td>
<td>.48</td>
</tr>
<tr>
<td>BL12</td>
<td>.33</td>
<td>.05</td>
<td>.37</td>
<td>.05</td>
<td>.00</td>
<td>.02</td>
<td>.00</td>
<td>.15</td>
<td>.41</td>
<td>.63</td>
</tr>
<tr>
<td>BL13</td>
<td>.35</td>
<td>.05</td>
<td>.40</td>
<td>.04</td>
<td>.00</td>
<td>.06</td>
<td>.02</td>
<td>.18</td>
<td>.30</td>
<td>.60</td>
</tr>
<tr>
<td>BL14</td>
<td>.26</td>
<td>.04</td>
<td>.29</td>
<td>.05</td>
<td>.00</td>
<td>.02</td>
<td>.00</td>
<td>.18</td>
<td>.45</td>
<td>.71</td>
</tr>
<tr>
<td>BL15</td>
<td>.32</td>
<td>.05</td>
<td>.37</td>
<td>.06</td>
<td>.00</td>
<td>.07</td>
<td>.02</td>
<td>.12</td>
<td>.36</td>
<td>.63</td>
</tr>
<tr>
<td>AVE-BL</td>
<td>.33</td>
<td>.06</td>
<td>.39</td>
<td>.05</td>
<td>.00</td>
<td>.04</td>
<td>.01</td>
<td>.12</td>
<td>.39</td>
<td>.61</td>
</tr>
<tr>
<td>DP1</td>
<td>.22</td>
<td>.11</td>
<td>.32</td>
<td>.07</td>
<td>.00</td>
<td>.09</td>
<td>.03</td>
<td>.02</td>
<td>.47</td>
<td>.68</td>
</tr>
<tr>
<td>DP2</td>
<td>.22</td>
<td>.11</td>
<td>.33</td>
<td>.05</td>
<td>.00</td>
<td>.02</td>
<td>.00</td>
<td>.06</td>
<td>.53</td>
<td>.67</td>
</tr>
<tr>
<td>DP3</td>
<td>.08</td>
<td>.04</td>
<td>.12</td>
<td>.04</td>
<td>.00</td>
<td>.05</td>
<td>.02</td>
<td>.25</td>
<td>.52</td>
<td>.88</td>
</tr>
<tr>
<td>DP4</td>
<td>.31</td>
<td>.15</td>
<td>.47</td>
<td>.05</td>
<td>.00</td>
<td>.06</td>
<td>.02</td>
<td>.03</td>
<td>.37</td>
<td>.53</td>
</tr>
<tr>
<td>DP5</td>
<td>.27</td>
<td>.05</td>
<td>.33</td>
<td>.07</td>
<td>.00</td>
<td>.09</td>
<td>.03</td>
<td>.07</td>
<td>.42</td>
<td>.67</td>
</tr>
<tr>
<td>DP6</td>
<td>.21</td>
<td>.04</td>
<td>.25</td>
<td>.06</td>
<td>.00</td>
<td>.02</td>
<td>.00</td>
<td>.09</td>
<td>.59</td>
<td>.75</td>
</tr>
<tr>
<td>DP7</td>
<td>.06</td>
<td>.01</td>
<td>.07</td>
<td>.05</td>
<td>.00</td>
<td>.06</td>
<td>.02</td>
<td>.30</td>
<td>.50</td>
<td>.93</td>
</tr>
<tr>
<td>DP8</td>
<td>.35</td>
<td>.07</td>
<td>.42</td>
<td>.05</td>
<td>.00</td>
<td>.06</td>
<td>.02</td>
<td>.13</td>
<td>.31</td>
<td>.58</td>
</tr>
<tr>
<td>DP9</td>
<td>.23</td>
<td>.06</td>
<td>.29</td>
<td>.08</td>
<td>.00</td>
<td>.10</td>
<td>.03</td>
<td>.11</td>
<td>.39</td>
<td>.71</td>
</tr>
<tr>
<td>DP10</td>
<td>.16</td>
<td>.04</td>
<td>.21</td>
<td>.06</td>
<td>.00</td>
<td>.02</td>
<td>.00</td>
<td>.10</td>
<td>.62</td>
<td>.79</td>
</tr>
<tr>
<td>DP11</td>
<td>.06</td>
<td>.02</td>
<td>.08</td>
<td>.05</td>
<td>.00</td>
<td>.06</td>
<td>.02</td>
<td>.29</td>
<td>.51</td>
<td>.92</td>
</tr>
<tr>
<td>DP12</td>
<td>.34</td>
<td>.08</td>
<td>.42</td>
<td>.05</td>
<td>.00</td>
<td>.07</td>
<td>.02</td>
<td>.05</td>
<td>.40</td>
<td>.58</td>
</tr>
<tr>
<td>AVE-DL</td>
<td>.21</td>
<td>.07</td>
<td>.28</td>
<td>.06</td>
<td>.00</td>
<td>.06</td>
<td>.02</td>
<td>.13</td>
<td>.47</td>
<td>.72</td>
</tr>
</tbody>
</table>

Notes: BLit (DPit) refers to the ith brand loyalty (deal proneness) item measured at time t, and AVE stands for average variance extracted.
it is possible to investigate the stability of brand loyalty and
defined. Specifically, we report comparisons with two other models: (1) a model in which the five brand loyalty and four
dead proneness items are averaged and the resulting unweighted composites are correlated and (2) model M2, in
which item-specific error is taken into account (as we mentioned previously, the fit of this model would probably be
ded acceptable by most researchers) and the correlations between brand loyalty and deal proneness are cor-
rected for attenuation using random error and item-specific error as an estimate of nonsubstantive variance (the conclu-
sions for separate two-factor models at each period are sub-
stantively the same).

Recall that the correlations between the occasion-specific
brand loyalty and deal proneness factors in the final model
were −.58, −.63, and −.65 (for Periods 1–3). In contrast, the
correlations between unweighted composites of the brand
loyalty and deal proneness items are −.22, −.16, and −.16.
When these correlations are corrected for attenuation, as in
model M2, the correlations increase, but the increase is
modest (i.e., −.29, −.22, and −.24), and the “true” correlation
is underestimated by a factor of 2 to 3. The reason for
the substantial bias in the estimated correlations is that the
conventional analysis ignores correlated errors of measure-
dment (due to item-subset and scalewide errors). These cor-
related errors of measurement are positive, whereas the sub-
stantive correlations are negative. If we remove the positive
error covariance from the observed measures and disattenu-
ate these adjusted observed correlations (which are −.36,
−.36, and −.37 in the current case), we obtain correlations of
−.58, −.63, and −.65 (as we reported previously).

If we also separate state variance from trait variance, the
correlation between trait brand loyalty and trait deal prone-
ness increases to −.70, because the correlation between the
state components of the two constructs is lower. In sum-
mary, after we separate extraneous sources of covariation
from both constructs and correct the observed correlations
with the proper index of reliability, the correlation between
the traits of brand loyalty and deal proneness increases (in
absolute value) from an average of −.25 in model M2 to
−.70, that is, by a factor of almost 3.

Measurement invariance and stability of brand loyalty
and deal proneness. As we show in Table 1, imposing
invariance constraints on the substantive and stable item-
specific factor loadings does not lead to a significant deteri-
oration in overall fit (M9) based on the fit indexes that take
into account model parsimony, and none of the Bonferroni-
adjusted modification indexes were significant. When we
also restricted the measurement intercepts to be invariant
(M10), the decrement in fit was more substantial, and two
intercepts needed to be freed (for the third deal proneness
item in Period 1 and the fifth brand loyalty item in
Period 1). The resulting model (M11) achieved the lowest
value of all models considered in terms of CAIC (1216),
and the other parsimony-based fit indexes were also as good
as in any other model: $\chi^2(325) = 760.09$, RMSEA = .026,
SRMR = .037, CFI = .981, and TLI = .979.

Because measurement invariance is adequately satisfied,
it is possible to investigate the stability of brand loyalty and
deal proneness (M12). Although we reject the hypothesis of
stability on the basis of a chi-square difference test
($\Delta\chi^2(4) = 21.99, p < .001$), this is primarily due to the large
sample size. Indeed, CAIC achieves its lowest value for this
model. The power of CAIC in model selection (Williams
and Holahan 1994) is supported by the finding that, prac-
tically speaking, the differences in the means of the occasion-
specific brand loyalty factors are less than .05 on a five-
point scale: 3.16 in 2000, 3.20 in 2002, and 3.16 in 2003 for
brand loyalty and 3.70 in 2000, 3.68 in 2002, and 3.71 in
2003 for deal proneness. These findings attest to a high
degree of stability for these two important marketing
constructs.

GENERAL DISCUSSION

Accurate measurement lies at the heart of marketing as a
science. Although the measurement of marketing constructs
has greatly improved in recent years, we believe that several
topics have received insufficient attention. Therefore, the
purpose of this article was to raise researchers’ awareness of
three key issues. First, in many areas of research, it is nec-
essary to draw a clear distinction between the trait and the
state aspects of a construct. The proposed measurement
model enables a sophisticated investigation of the stable
and transient components of a construct and avoids the prob-
lems associated with assessments of stability through test–
retest correlations. Second, our extended paradigm for scale
development explicitly accounts for various sources of
measurement error and enables a much more detailed diag-
osis of the quality of individual items and the measure-
ment scale as a whole. Furthermore, as we show in the
empirical illustration, it effectively accounts for the biasing
effects of random and systematic measurement error, which
can be extremely damaging to the validity of the substantive
conclusions derived from the research. Third, the procedure
incorporates the item and scale means into the measurement
analysis, makes assessments of measurement invariance an
explicit component of the scale validation process, and ulti-
mately leads to cross-sectional and longitudinal compari-
sions of means (as well as relationships between constructs)
that are methodologically justified.

It is apparent that the proposed methodology requires
data that are more effortful to obtain than those used in con-
ventional procedures. That is, large samples of respondents
(typically >500, as we discuss in more detail subsequently)
must be surveyed at least three times, though one-year time
intervals between successive measurements for trait versus
state assessment are not necessarily needed; periods as
short as two to four weeks have often been used (see Bear-
den and Netemeyer 1999). This is the price researchers
must pay for being able to submit a scale to a stringent con-
struct validation process and derive diagnostic information
from the analysis. Our empirical illustration shows that this
is worth the additional effort.

First, we found that about one quarter of the total vari-
ance was systematic (nonrandom) measurement error, most
of it stable over time. This is a matter of concern, particu-
larly because these scales are used in marketing practice
(the data are collected annually by a large market research
agency).

Second, scale reliability using the conventional proce-
dure was substantially overestimated. This result is proba-
bly not unique to the current context because the sources of
systematic error will also be present in other scales. A con-
sequence of this finding is that it may require a rethinking
of conventional reliability standards (Nunnally 1978),
which are based on Equation 1 or simplifications thereof (e.g., Cronbach’s alpha).

Third, we found that ignoring systematic measurement error had a substantial biasing effect on the results because the correlation between the substantive constructs of brand loyalty and deal proneness was underestimated by a factor of almost 3 (–.25 versus –.70). In general, the direction of the bias depends on the sign of the true correlation. Measurement errors are typically positively correlated. If the true correlation between two constructs is negative, it will be underestimated, as in our illustration. If the true correlation is positive, it will be overestimated.

It is not our intention to imply that researchers should always collect longitudinal data. However, the required data collection effort might be an important element in research in which the primary contribution is to develop a measurement instrument for a construct. The additional diagnostic power derived from applying the procedure described in our article might increase the confidence of other researchers in the new scale and may prevent the need for subsequent scale modifications and testing. Although the requirement of three repeated measurements may seem daunting, we believe that there are at least two opportunities for collecting such data.

First and foremost, consumer panels organized by market research agencies and universities (particularly Internet panels) are natural candidates for repeated surveys. They make the data collection less burdensome, guarantee a reasonable response rate, and, as an added bonus, make the results more generalizable because of the use of a heterogeneous participant pool. Even when the initial scale development does not include repeated measurements, researchers with access to consumer panels can conduct validity studies in which commonly used instruments are subjected to a rigorous measurement analysis. Second, participants in a subject pool or students enrolled in a class can be surveyed repeatedly throughout the semester or even during the course of the year. For example, the data used in the current application consist of nine items, and it should not take more than ten minutes to complete the questionnaire on a given occasion.

Further Research

The models we discussed in this article tend to be highly parameterized, so it is beneficial to have large samples because estimation problems, such as nonconvergence, are less likely in that case. Our sample was large, and no estimation problems occurred. Preliminary simulation work that we conducted shows that for the type of model considered in this article, convergence rates are high for a sample size of 2000 (97%) and good for a sample size of 1000 (88%), but they drop off at smaller sample sizes of 500 and especially 250 (69% and 43%, respectively). However, even the lower convergence rates compare favorably with those reported for confirmatory factor analyses of multitrait-multimethod data (e.g., Marsh and Bailey 1991). Further research should investigate this issue in much greater detail.

The proposed methodology can also be extended in various ways. Antecedents and consequences of the focal constructs can be included in the specification, and structural relationships among the substantive constructs can be investigated. Trajectories of scores over time can be studied at the individual level, and hypotheses about systematic influences on individual change processes can be tested (e.g., Tisak and Tisak 2000). Known participant heterogeneity can be accommodated by estimating a multigroup, over-time model in which the groups are defined a priori (e.g., men versus women). Unknown participant heterogeneity may be investigated by specifying mixture models that allow for a limited number of classes of respondents with different model parameters (Jedidi, Ragpal, and DeSarbo 1997).

Possibly the most important avenue for further research we suggest is the need to understand the sources of error that contribute to observed scale scores and the psychological mechanisms that are responsible for these errors (for a recent discussion, see Podsakoff et al. 2003). Although previous research has shown that measurement error frequently accounts for substantial amounts (if not the majority) of observed variance in scale scores (e.g., Cote and Buckley 1987), it is often treated as a mysterious amalgam of unknown influences on people’s responses to questionnaire items. We have tried to provide a structure for distinguishing among different types of error. However, to manage the various error types, a lot more needs to be discovered about what gives rise to these errors. Understanding the causes of the various types of error is a necessary step in the development of procedures to reduce error in marketing measurement, which is crucial for the further growth of marketing as a science. Hopefully, researchers will take our lead and make measurement error the substantive focus of some of their work.
Appendix

LISREL COMMAND FILE FOR MODEL M8

DA NI=27 NO=0
LA
dp11  dp21  dp31  dp41  bl11  bl21  bl31  bl41  bl51
dp12  dp22  dp32  dp42  bl12  bl22  bl32  bl42  bl52
dp13  dp23  dp33  dp43  bl13  bl23  bl33  bl43  bl53
RA F ile[directory where data file is located]
MO NY=27 NE=21 NK=2 LY=FI TE=DL,FR TY=FR GA=FI PS=SY,FI PH=FR
LE
DPT1 BLT1 DPT3 BLT3 DPT4 BLT4
SIE1 SIE2 SIE3 SIE4 SIE5 SIE6 SIE7 SIE8 SIE9
SSEPos TSEPos1 TSEPos2 TSEPos3 SSENeg SSWE
LK
DP
BL
FR LY 2 1 LY 3 1 LY 4 1 LY 6 2 LY 7 2 LY 8 2 LY 9 2
FR LY 11 3 LY 12 3 LY 13 3 LY 15 4 LY 16 4 LY 17 4 LY 18 4
FR LY 20 5 LY 21 5 LY 22 5 LY 24 6 LY 25 6 LY 26 6 LY 27 6
FR LY 10 7 LY 11 8 LY 12 9 LY 13 10 LY 14 11 LY 15 12 LY 16 13
FR LY 17 14 LY 18 15 LY 19 7 LY 20 8 LY 21 9 LY 22 10 LY 23 11
FR LY 24 12 LY 25 13 LY 26 14 LY 27 15
MA LY
1 0 0 0 0 0 0 1 0 0 0 0 0 0 0 1 1 0 0 0 1
1 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 1 1
1 0 0 0 0 0 0 0 1 0 0 0 0 0 0 1 1 0 0 0 1
1 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 1 1
0 1 0 0 0 0 0 0 0 0 0 0 0 0 1 0 1 0 0 0 1
0 1 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 1 1
0 1 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 1 1
0 1 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 1 1
0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 1
0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1
0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 1
0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1
0 0 0 0 1 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 1
0 0 0 0 1 0 0 0 0 0 0 0 0 0 1 1 0 0 0 0 1
0 0 0 0 1 0 0 0 0 0 0 0 0 0 1 0 1 0 0 0 1
0 0 0 0 1 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 1
0 0 0 0 1 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 1
0 0 0 0 1 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 1
0 0 0 0 1 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 1
0 0 0 0 1 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 1
0 0 0 0 1 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 1
0 0 0 0 1 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 1
MA TE
[starting values for TE matrix]
MA TY
[starting values for TY matrix]
FR GA 3 1 GA 4 2 GA 5 1 GA 6 2
VA 1 GA 1 1 GA 2 2
MA GA
[starting values for GA matrix]
FR PS 1 1 PS 2 2 PS 3 3 PS 4 4 PS 5 5 PS 6 6 PS 7 7 PS 8 8
FR PS 9 9 PS 10 10 PS 11 11 PS 12 12 PS 13 13 PS 14 14
FR PS 15 15 PS 16 16 PS 17 17 PS 18 18 PS 19 19 PS 20 20
FR PS 21 21 PS 2 2 PS 4 3 PS 5 6 5
MA PS
[starting values for PS matrix]
MA PH
[starting values for PH matrix]
OU RS SC MI NS ND=3 ad=OFF WP
REFERENCES


