Marsh, H. W, Lüdtke,O., Nagengast, B., Morin, A. J. S., Trautwein, U., Von Davier, M.. (2012). Why Item Parcels Are (almost) Never Appropriate: Two Wrongs Do Not Make A Right – Camouflaging Misspecification with Item-Parcels in CFA Models

Why Item Parcels Are (almost) Never Appropriate: Two Wrongs Do Not Make A Right—Camouflaging Misspecification with Item Parcels in CFA Models

Herbert W. Marsh [1,2,3] Oliver Lüdtke [4]; Benjamin Nagengast [2, 5];

Alexandre J. S. Morin [1]; Matthias Von Davier [6]

[1] University of Western Sydney; [2] University of Oxford, UK; [3] King Saud University, Saudi Arabia; [4] Humboldt University, Germany; [5] University of Tübingen, Germany; [6] Educational Testing Service, USA

8 October 2011

Revised 20 September 2012

Revised 10 March 2013

Author note

The authors would like to thank Alexander Robitzsch, Tihomir Asparouhov, Bengt Muthén and Cameron McIntosh for helpful comments at earlier stages of this research. This research was supported in part by a grant to the first author from the UK Economic and Social Research Council. Requests for further information about this investigation should be sent to Professor Herbert W. Marsh, University of Western Sydney; E-mail: H.Marsh@uws.edu.au

Abstract

The present investigation has dual a focus: to evaluate problematic practice in the use of item parcels and to suggest exploratory structural equation models (ESEM) as a viable alternative to the typical independent clusters confirmatory factor analysis model (ICM-CFA; with no cross-loadings, subsidiary factors or correlated uniquenesses). Typically it is ill-advised to: (a) use item parcels when ICM-CFA models do not fit the data; and (b) retain ICM-CFA models when items cross-load on multiple factors. However, the combined use of (a) and (b) is widespread, and often provides such misleadingly good fit indexes that applied researchers might believe that misspecification problems are resolved—that two wrongs really do make a right. Taking a pragmatist perspective, in four studies we demonstrate with responses to the Rosenberg Self-Esteem Inventory, big-five personality factors, and simulated data that even small cross-loadings seriously distort relations among ICM-CFA constructs or even decisions on the number of factors; although obvious in item-level analyses, this is camouflaged by the use of parcels. ESEMs provide a viable alternative to ICM-CFAs and a test for the appropriateness of parcels. The use of parcels with an ICM-CFA model is most justifiable when the fit of both ICM-CFA and ESEM models is acceptable and equally good, and substantively important interpretations are similar. However, if the ESEM model fits the data better than the ICM-CFA model, then the use of parcels with an ICM-CFA model typically is ill-advised—particularly in studies that are also interested in scale development, latent means and measurement invariance.

Keywords: Item parcels; exploratory structural equation models; misfit; measurement invariance; scale development.

The present investigation has a dual purpose in relation to critical measurement issues that face applied researchers. The first and primary purpose is to explore potentially serious limitations in typical current practice in the use and misuse of item parcels in factor analysis and structural equation models (SEMs). Item parcels are the sum or mean of responses to several indicators designed to measure the same construct. Thus, analyses are conducted on a smaller number of parcels designed to measure each factor rather than on a larger number of items (see subsequent discussion). Yang, Nay and Hoyle (2010) argue that item parceling is the prevailing approach for including scales with many items in factor analysis and SEM models. We intentionally take an extreme (empirical pragmatic) position: that the use of item parcels is (almost) never appropriate a priori; we offer logical and empirical support for this extreme position, and then explore pragmatic instances in which this extreme position might be qualified. For purposes of this article we focus primarily on the evaluation of factor loadings, measurement error, and relations among constructs where the use of parcels is most common. However, we also emphasize that if the focus is scale construction or on latent means, measurement invariance, and differential item functioning – or even analyses needed to justify the interpretation of manifest means – then the a priori use of item parcels is particularly dubious (e.g., Bandalos & Finney, 2001).

The second, subsidiary purpose is to compare the use of exploratory structural equation modeling (ESEM) used in a confirmatory manner and the traditional independent clusters confirmatory factor analysis model (ICM-CFA; with no cross-loadings, secondary factors or correlated uniquenesses). We argue that the two issues are closely related in that it is precisely when ESEM outperforms ICM-CFA that the use of item parcels is most fraught with potential problems. Although not yet widely used in applied research, ESEM integrates many of the advantages of EFA and CFA.

In the present investigation, we explore these issues with responses derived from the Rosenberg Self-Esteem Inventory (Study 1) and to Neuroticism and Extraversion (two big-five personality factors; Study 2), and with simulated data with varying approximations to unidimensionality (Studies 3 and 4). However, the conceptual and methodological concerns have important implications for all disciplines of psychology (and the social sciences more generally) that rely on factor analysis.

The over-arching rationale for this study is the now widely accepted recognition that statistical models—including factor analysis and SEMs—only reflect approximations to reality that are always

wrong (e.g., Cudeck & Henly, 1991; MacCallum, 2003; McDonald, 2010; Marsh, Balla & McDonald, 1988; Thissen, 2001; but also see Box, 1979; Tukey, 1961; Thurstone, 1930). As emphasized by MacCallum (2003, p. 114) in his Presidential Address:

> *Regardless of their form or function, or the area in which they are used, it is safe to say that these models all have one thing in common: They are all wrong. Simply put, our models are implausible if taken as exact or literal representations of real world phenomena.*

From this perspective, it is essential for applied researchers to evaluate how model misspecification influences their interpretations and conclusions. Here we demonstrate the importance of this issue in relation to current practices of using item parcels in combination with ICM-CFA models. Fundamentally, the ICM-CFA model is based on the assumption of the unidimensionality of all factors included in the model. However, in the same way that all statistical models are wrong, from a philosophical perspective we claim that when applied to real data, this assumption of unidimensionality is always wrong. From an empirical perspective the assumption of undimensionality for unsaturated models will always be rejected for a sufficiently large N. Although all models are wrong, some are more useful than others (Box, 1979). Hence, the critical issues are how badly this undimensionality assumption is violated, how serious the violation has to be to undermine the usefulness of the model, and the extent to which critical interpretations of the data are affected by this misspecification. Our overarching message in this situation—as well as in applied research more generally—is that it is better to evaluate potential sources of misspecification systematically than to ignore or camouflage them.

The Number of Indicators and The Use of Item Parcels

Is More Ever too much—the Number of Indicators per Factor

How many indicators of each factor are needed? Marsh et al. (also see De Winter, Dodou & Wieringa, 2009) argued that “more is never too much” for the number of indicators as well as the number of participants. For simulated data based on a population-generating model that approximates the unidimensionality assumption of ICM-CFA, it might be reasonable to have only a few indicators per factor (e.g., enough to model and control measurement error). However, real data rarely have these ideal properties. Mostly, applied researchers work with factors that are hypothetical constructs and have to be validated in relation to a construct validity approach. For real data, unidimensionality and pure indicators

are an ideal to strive towards (i.e., a convenient fiction), but are rarely if ever achieved. As noted by MacCallum (2003, p. 134): *Studies based on the assumption that models are correct in the population are of limited value to substantive researchers who wish to use the models in empirical research*. Thus, almost all indicators are factorially complex if actually put to the test and often include a combination of true-score variance, random variance, specific variance, correlated uniquenesses, etc. In studies based on multidimensional constructs or a variety of related factors, most indicators will cross-load on at least some of the other factors (if allowed to do so). Most random error only appears to be random because applied researchers have not looked closely enough for sources of systematic error variance. Most indicators will also have some method effects related to the way data was collected, the idiosyncratic wording of items, response biases, etc. Most of these complexities with real data might also influence or bias relations with other constructs used to assess construct validity.

Historically it was common to have 10–15 or more items per scale for the most widely used psychological tests, and standardized achievement tests typically have considerably more than 15 items. In order to enhance the generalizability of constructs, it is better to have more indicators (Marsh et al., 1998). At least implicitly, tests are typically constructed under the assumption that the available indicators are a subset of a potentially very large (or infinite) number of indicators of the same construct (McDonald, 2010). Except under unrealistic assumptions of pure unidimensionality, constructs measured with only a few indicators are likely to be "bloated specifics" and might not really represent the label given to the factor (see Little, et al., 1999). In the same way that it is always better to have larger sample sizes in relation to participants, it is also better to have more independently answered items. Indeed, it is interesting to explore the logic of the argument for having only a few items when applied to the number of participants. In a perfect (simulated) data world, it might be possible to find "truth" based on only a few participants. However, with real data this would seriously undermine the generalizability of the findings. The same argument applies to making generalizations based on a single or only a few indicators of most constructs. Under rare circumstances, it might be possible to argue that only responses based on a few participants or a few indictors are needed, but typically this undermines the generalizability of the interpretations.

Item Parcels

As explained above, it is better to have more indicators per factor. However, there is an understandable reluctance on the part of applied researchers to incorporate large numbers of indicators—particularly in large, complex models that involve many different factors. Indeed, there can be technical problems in estimation and even identification of factor models when the number of estimated parameters and measured variables is too large relative to the sample size. Hence, applied researchers have sought a judicious compromise between parsimony and accuracy. One widely employed compromise (see Marsh, et al. 1988) is to collect many items, but to use item parcels in the analyses. For example, in a psychological instrument assessing 10 factors with 12 items each, the 120 items could be used to form three four-item parcels for each factor that are actually used in the analysis (e.g., take the average of the 1st, 5th, and 9th indicators to form one parcel, the 2nd, 6th, and 10th indicators to form a second parcel, and so forth). It might be claimed that this would result in a much more parsimonious model based on 30 (parcel) indicators rather than 120 (item) indicators, substantially reducing the number of measured variables and parameter estimates. However, model parsimony cannot really be compared for item- and parcel-indicators, as the formation of parcels defines a new set of variables and changes the nature of the data. More generally, the use of item parcels is fraught with difficulties (e.g., how to justify which items go into each parcel given the huge number of possibilities) and should be pursued with caution.

Historical Perspective. The analysis of item parcels instead of items has a long history in psychology, dating back at least to Cattell (1956, 1974) and currently is widely used in CFA studies (e.g., Little, Cunningham, Shahar, & Widaman, 2002). The posited advantages of using a relatively few parcels per factor rather than a larger number of items include: increased reliability of item-parcel responses; better approximations of normality assumptions and normal theory-based estimation; improved goodness of fit; fewer parameters to be estimated (optimizing the ratio of sample size to the number of measured variables or estimated parameters); more stable parameter estimates; reduction in idiosyncratic characteristics of items; and simplification of model interpretation (see Bandalos & Finney, 2002; Little et al., 2002; Marsh & O'Neill, 1984; Marsh et al., 1998; Williams & O'Boyle, 2008; Yang et al., 2010).

In a review of articles from major education, psychology and marketing journals, Bandalos and Finney (2001; Bandalos, 2002) found that 20% of the applied SEM/CFA studies used some type of parceling procedure. Furthermore, the number would be much higher if they had included studies based on

scale scores (a single parcel based on a weighted or unweighted average of item responses; Williams & O'Boyle, 2008). However, Bandalos and Finney further noted that even though earlier research by Marsh (e.g., Marsh & O'Neill, 1984) was the most frequently cited reference in support of parceling (70% of studies using parceling in their review cited Marsh's research), many researchers failed to heed Marsh's advice that items being parceled should be reasonably unidimensional (i.e., an ICM-CFA model should be able to fit the item-level data). Following from Marsh and O'Neill (1984), Bandalos and Finney (2001; Bandalos, 2002), and many others, most reviews of parceling and many studies that use parceling give at least token mention this assumption of unidimensionality, but it is rarely evaluated systematically for the models and data under consideration (Williams & O'Boyle, 2008).

Reviews of the use of parceling strategies (e.g., Bagozzi & Edwards, 1998; Bandalos & Finney, 2001; Little, et al., 2002; Marsh & O'Neill, 1994; Marsh et al., 1998; Sass & Smith, 2006; Sterba and MacCallum; Williams & O'Boyle, 2008) have been generally positive about parceling under appropriate conditions: when the focus is on relations between constructs (i.e., factor correlations or path coefficients) rather than scale development, and there is good a priori information to support the posited factor structure such that each item loads on one and only one factor (i.e., there are no cross-loadings) with no correlated uniquenesses, and no secondary factors—in short, when an ICM-CFA model fits the data at the item level. Indeed, it is already known that violations of unidimensionality can be confounded with common factor variance/covariance with item parceling as shown in the context of Sterba and MacCallum's (2010) theoretical framework which extends MacCallum, Widaman, Zhang & Hong (1999), MacCallum & Tucker (1991), and simulation studies (e.g. Bandalos, 2002; Hall et al., 1999; Sterba, 2011). However, Sterba and MacCallum (2010) also emphasized that even when appropriate conditions are met in the population, large sampling variability in samples led to substantial variation in the results associated with different parcel allocations. Furthermore, Williams and O'Boyle (2008) indicated that in practice, applied researchers frequently do not explicitly test the dimensionality of their constructs, and some use parceling even though empirical tests show that assumptions of unidimensionality are violated. Interestingly, they noted that a few studies tested multidimensionality with preliminary exploratory factor analyses (e.g., Anderson, 2002; also see Coffman & MacCallum, 2005)—an approach particularly relevant to the present investigation. However, even when tests of unidimensionality are explicit, they are often conducted separately for each

factor, thus ignoring the likely problem of items cross-loading on more than one factor that violates the appropriate use of item parcels.

In summary, like many others before us (e.g., e.g. Bandalos & Finney 2001; Bandalos, 2002; Hall, Snell & Foust, 1999; Kishton & Widaman, 1994; Marsh & O'Neill, 1984), we emphasize that the applied researcher must establish unidimensionality before parceling. Although not a new message, we contend that it is largely ignored in applied research. It is not sufficient to merely note this as a limitation in applied research without empirically evaluating its consequences. Whereas citing previous research in support of unidimensionality of key measures in one's study (e.g., Bandalos & Finney, 2001) is clearly useful and appropriate, it is not sufficient justification for the assumption in a new study. Indeed, it is likely that a factor shown to be relatively unidimensional on its own or in combination with one set of factors will not be unidimensional (e.g., there are cross-loadings) when considered in the context of different factors, so that tests of unidimensionality are typically idiosyncratic to particular studies. Hence applied researchers should base tests of unidimensionality on the models actually applied to their own data.

Item parceling strategies. Sterba and MacCallum (2010) emphasize that even when the conservative unidimensionality requirements for the use of item parcels are met, there can be substantial variability in parameter estimates and fit indexes, due to different allocations of items to parcels; noting that it is possible to construct three 4-item parcels from a 12-item scale in 34,650 different ways. They demonstrated this potential problem with simulated data and with real data consisting of item parcels based on big-five personality responses for which they reported parceling had been used extensively.

Bandalos (2008) noted that in applied research, item parcels typically are formed in ad hoc ways. However, researchers have explored different systematic approaches to constructing parcels (Williams & O'Boyle, 2008; also see Hall, Snell & Founst, 1999; Little, et al., 2002; Rogers & Schmitt, 2004) including random assignment of items to parcels, systematically distributing items with known characteristics (e.g., positive and item wording; correlations among items, correlated uniquenesses among items; item difficulty, discriminability, or factor loadings based on preliminary analyses at the item level) across different parcels, or systematically placing items with known characteristics within the same parcel.

For present purposes we distinguish between ***homogeneous parceling*** and ***distributed parceling strategies*** (also see Bandalos, 2008)**.** In homogeneous parceling strategies, closely related items are placed

in the same parcel – items that share common characteristics (e.g., item wording), systematic variation, common method effects or, perhaps, items that are from the same facet or subfactor of the overarching factor. In distributed parceling strategies, items that share a source of systematic variation are distributed across different parcels either randomly or systematically. Thus, many approaches to forming parcels (e.g., random assignment) are implicit (or de facto) distributive strategies even if this is not the explicit intent of the applied researcher.

For us, the key distinction between these strategies is that homogeneous strategies highlight potential sources of misfit whereas distributive strategies confound, obscure, or mask sources of misfit with the allocation of items to factors so that the misfit is camouflaged and seems to 'disappear without a trace'. For example, in relation to positively and negatively worded items designed to measure the same factor, a homogeneous strategy would construct parcels that are homogeneous in relation to item wording, thus highlighting this effect if it was a source of misfit. Conversely, an explicit distributive strategy would construct item parcels with approximately equal numbers of positively and negatively worded items and an implicit distributive strategy might randomly assign items to parcels such that by chance most parcels would contain a mix of positively and negatively worded items. For each of the distributive strategies, the item-wording effect would 'disappear' in that it was completely confounded by the construction of item parcels. In their review, Little et al. (2002; also see Kishton & Widaman, 1994) noted that the homogeneous strategy can result in problems (e.g., unstable solutions or unacceptable parameter estimates), whereas the distributive strategy was less prone to these problems. Although they recommended the distributive strategy, they conceded that part of what they called 'compelling evidence' for this approach was that it confounded sources of misfit. Coffman and MacCallum (2005) also compared these strategies, but concluded that "how the parcels are constructed is less important than the fact that they are used" (p. 253).

In evaluating these different approaches to parceling it is important to note that if the constructs are truly unidimensional—a prerequisite for the appropriate use of parcels—then each of these strategies should yield reasonably similar results. This might explain why some researchers found little difference between the homogeneous and distributive strategies—particularly when based on simulated data that met the assumption of unidimensionality. However, if the unidimensionality assumption is violated seriously,

this misspecification in the ICM-CFA model is likely to be camouflaged by the distributive strategy to a much greater extent than with the homogeneous strategy. Indeed, the underlying rationale of the explicit distributive strategy only makes sense when the assumption of unidimensionality is violated, thus precluding the appropriate use of item parcels.

Apparent Support for the use of item parcels. Simulation studies in support of the use of item parcels typically begin with data that at least approximately meet the unidimensionality requirement for the appropriate use of parcels (e.g., Alhija & Wisenbaker, 2009; Marsh, Hau, et al., 1998; Sass & Smith, 2006; Yang, et al., 2010) and tend to be cautiously positive about their usefulness. For example, Alhija and Wisenbaker (2009) concluded that item parceling had negligible effects on parameter bias and standard errors. Also, when studies claim that item parceling does or does not affect fit, estimates, or standard errors, it is important to note what is being compared (e.g., parcel and item solutions, alternative parceling strategies, variability of parcel solutions) under what conditions. For example, using simulated data where undimensionality held in the population generating model, Hall et al. (1999; Study 1) and Sass and Smith (2006; Study 1) found only small differences due to parceling allocation when sampling variability was low (e.g., large N) but Sterba and MacCallum (2010) found meaningfully large differences under conditions of moderate or high sampling error. However, when ideal conditions do not prevail (i.e., when ICM-CFA is not appropriate), the use of item parcels tends to hide possibly trivial or potentially serious problems of misfit. In applied research with real data, this will usually be the case, as the ideal ICM-CFA model is only intended to provide a parsimonious approximation to the complexity of real data. Furthermore, unless these ideal conditions do hold, the fit of multiple factors based on a reasonably large number of items will usually be substantially poorer than the corresponding fit based on a much smaller number of parcels constructed from the items. However, it is important to stress that this does not mean that the parcel model is better: First, this model was based on a different set of variables, so it was no longer a model for the original data, and second, by forming new variables (parcels), the sources of misfit are camouflaged by the use of parcels, with resulting biases in the parameter estimates.

The use of parcels tends to camouflage potentially serious violations of unidimensionality and of the ICM-CFA model. For example, assume that 12 items designed to reflect a single underlying factor actually reflect three different factors or, perhaps, three subfacets of a single factor. An ICM-CFA solution

positing a single factor based on item responses will probably provide an unacceptable fit that appropriately represents the violation of unidimensionality assumptions. If, however, four parcels are constructed such that each parcel has one indicator from each of the underlying three factors (an explicit distributed parceling strategy), the single-factor solution is likely to provide an apparently good fit, and to inappropriately support the construct's unidimensionality and the erroneous single-factor a priori assumption of the researcher. Similar arguments can be made when the use of parcels camouflages potential cross-loadings in the typical ICM-CFA structure based on multiple factors. Although the consequences of minor cross-loadings for items representing different factors (or other violations of unidimensionality) may be substantively trivial, this is not always the case. Whereas it can be argued that such violations of unidimensionality are substantively unimportant, such arguments should be made explicit and tested empirically, rather than being camouflaged by the use of parcels.

Little et al. (2002; also see Coffman & MacCallum, 2005; Williams & O'Boyle, 2008) argued for a pragmatist perspective in support of the use of item parcels. They suggested that if the focus of research is on relations among constructs, then eliminating cross-loadings or correlated uniquenesses with item parceling is as effective as explicitly modeling these effects. However, implicit in this suggestion is the typically untested – but usually erroneous – assumption that relations among constructs are accurately represented if cross-loadings or correlated uniquenesses are eliminated by parceling—a major focus of the present investigation. This issue is somewhat analogous to interpretations based on a structural equation model (relations among constructs) without fully evaluating the measurements model (McDonald, 2010). Recognizing this potential problem, Little et al. (2002, p. 169) conceded that "if parcels can obscure invalid assumptions in a model, then a researcher can never be assured that a proposed model is legitimate or correctly specified". However, it is our contention that this will usually be the case so that analyses based on parcels can almost never be trusted a priori without further empirical tests. An extreme pragmatist (ostrich) perspective might be to simply ignore or deny any problems associated with the use of parcels. Conversely, Little's pragmatist perspective is not as extreme and recognizes that the use of parcels is based on assumptions that can render the practice as counter-productive. Here we advocate an alternative, 'empirical pragmatist' approach to the appropriate use of parcels; empirical tests of the assumptions

underlying the use of parcels and an evaluation of consequences of violations of these assumptions in terms of pragmatic interpretations in relation to the purposes of the study.

Similarly, Bandalos (2008) proposed that the widely cited improved fit produced by item parceling might be the result of masking (or camouflaging) rather than correcting the source of model misfit. Thus "the efficacy of item parceling depends on the unidimensionality of the items being combined. When this assumption is not met, the use of parcels can obscure rather than clarify the factor structure of the data and result in biased parameter estimates and fit index values" (Bandalos, 2008, p. 212). Pointing out that even in CFA studies there is often ambiguity in the appropriate number of factors, Bandalos (2002, 2008) suggested that even testing the correct number of factors to be considered might be masked by the use of parcels. Indeed, she noted that the problem of underfactoring has been studied extensively and that, even when there is a clear a priori basis for the number of factors, these assumptions might not be accurate. She explored this issue by evaluating misspecified models in which two separate factors were modeled as a single factor. She argued that this problem is relevant to the situation in which the applied researcher incorrectly models too few factors (an under-factoring issue) and unknowingly constructs parcels from items from two different factors. Although this misspecification was obvious at the item level, it was not so apparent in parcel solutions. Particularly for parcels constructed using distributed strategies, the separation of the different factors is confounded with the formation of factors so that the parcel solution typically fails to identify this misspecification. More importantly, this approach results in substantial bias in the estimates of relations among factors. Indeed, all of the parceling strategies considered in her study of misspecified models resulted in biased relations among constructs, ranging from 20% to over 130%. However, the distributive parceling strategies resulted in the best fit (camouflaging the misfit) and the most biased parameter estimates. On the basis of her study, Bandalos questioned the usefulness of parceling altogether and suggested alternative approaches to evaluate the factor structure at the item level more effectively.

Exploratory Structural Equation Models (ESEM).

CFA tests of a priori factor structures are typically based on an independent cluster model (ICM) in which each item loads on a single factor. Although there are many methodological and strategic advantages to ICM-CFAs, Marsh (2007; Marsh, Hau, & Grayson, 2005) argued that few multidimensional assessment instruments met even minimal standards of goodness of fit based on ICM-CFAs. Here we argue that part of

the problem is undue reliance on overly restrictive ICM-CFA model. Marsh et al. (2009, 2010, 2011a, 2011b) argued that this failure to achieve acceptable levels of fit has led to many compensatory strategies that might in some instances be dubious, counterproductive, misleading, or simply wrong—including the use of item parcels that is the focus of the present investigation. Furthermore, the misspecification of factor cross-loadings (constraining them to be zero when they are not) usually results in inflated factor correlations that might lead to biased estimates in structural equation models incorporating other outcome variables.

As an alternative to the ICM-CFA approach, Marsh et al. (2009a, 2009b, 2011a, 2011b) proposed ESEM, which provides an integration of many of the best aspects of CFA, SEM, and traditional EFAs. The ESEM approach differs from the typical CFA approach in that all factor loadings are estimated, subject to constraints necessary for identification (for further details of the ESEM approach and identification issues, see Asparouhov & Muthén, 2009; Marsh et al., 2009, 2010, 2011a, 2011b). For very simple factor analysis models (with no correlated uniquenesses, no invariance or additional parameter constraints, no structural paths, and based on a single wave of data for a single group with no multilevel structure), ESEM models are EFA models. In fact, the ESEM framework simply merges EFA, CFA, and SEM into an integrated and overarching framework, adding flexibility to all of its subcomponents yet still allowing for the estimation of pure EFA, CFA, or SEM models. Here, we refer to ESEM models to describe models in which at least some of the factors are defined as EFA factors and contrast this to ICM-CFA models where all factors are based on an ICM-CFA model. Of course, many models that applied researchers want to estimate and contrast with CFA or SEM models – including those in the present investigation – cannot be fit with a traditional EFA approach so that the ESEM model is much more flexible than traditional EFA models. For example, for longitudinal data the ESEM approach can evaluate the factor structure separately at each wave of data or with all waves combined in the same model and test measurement invariance over time – using the same rationale as longitudinal CFAs. Similarly, like CFA, ESEM can be used to test full measurement invariance across multiple groups, whereas this is not easily accomplished with EFA. In this sense, traditional EFAs are merely a special case of the more general ESEM framework, as CFAs, regression and SEMs are all part of an over-arching SEM framework. Indeed we also think of CFA and

SEM as special cases of the more general ESEM in that CFA and SEM models are nested under corresponding ESEM models (see Morin, Marsh, Nagengast, 2012), thus facilitating their comparison.

For present purposes we propose that the comparison of ESEMs used in a confirmatory manner and the traditional ICM-CFA provides a potentially useful test of the minimal conditions under which it might be appropriate to use item parcels. In particular, if the ESEM solution fits the data better than the ICM-CFA solution and results in substantively different interpretations of the data, then the assumptions underlying the use of item parcels are violated and their use is fraught with potential problems.

The Present Investigation

Issues related to the appropriateness of ICM-CFAs and item parcels have wide applicability for all disciplines of psychology (and the social sciences more generally) that use factor analysis. In particular, as emphasized by Marsh (2007; Marsh, Hau & Grayson, 2005; also see Cook, Kallen, & Amtmann, 2009; Hopwood & Donnellan, 2010; Reise, 2012), there are apparently few widely used psychological measures assessing multiple factors with a reasonable number of items per factor that are able to meet current standards of fit based on an ICM-CFA model. In such cases, the basic unidimensionality assumption required for the appropriate use of item parcels is not met. The inappropriate use of item parcels in this situation typically camouflages misfit without resolving the potentially serious bias in parameter estimates, and results in substantially inflated goodness of fit. Following McDonald (2010) and others, we emphasize that part of the evaluation of a factor structure should be a careful evaluation of parameter estimates, residuals, and modification indices at the item level, which cannot be pursued appropriately if applied researchers rely solely on the use of item parcels.

For us the fundamental justification of analyses of parcels is a common belief that when the use of item parcels is appropriate, parcel solutions are better behaved than item solutions (e.g., Little, et al., 2002). However, support for this assumption is surprisingly ephemeral. In a test of the assumption, Marsh et al. (1998) conducted a Monte Carlo study in which they compared CFAs based on either two, three, four, six, or 12 items per latent construct and two, three, four, or six parcels created from all 12 items per latent construct. The results clearly demonstrated the more-is-better principle in that constructs defined with more items were better defined than solutions based on fewer items. Marsh et al argued for this conclusion on the basis of domain sampling theory (e.g., Nunally, 1978; Little, et al., 1999) and data

aggregation principles (e.g., Rushton, Brainerd, & Pressley, 1983), as well as their empirical results. Little et al. (2002) similarly concluded that "a general conclusion that can be drawn from the foregoing psychometric considerations is that the use of additional items yields a more encompassing and inclusive representation of a construct" (p. 158). However, more interesting for the present article is whether given a relatively large number of items per factor (12 in that study), is it better to conduct analyses based on items or parcels, and whether this decision depends on the sample size. Importantly, the results showed that item solutions based on 12 items per factor performed as well or better than parcel solutions constructed from the 12 items even when N was very small. Commenting specifically on this aspect of the study, Little et al. (2002) claimed that "a proponent of parceling, however, may counter that the simulation was not a fair" (p. 162) because the data were too clean (i.e., based on an ICM-CFA population-generating model). However, we would argue for exactly the opposite interpretation; that parceling is only appropriate when the data provide a reasonable approximation to the ICM-CFA model, so that the earlier Marsh et al. study was biased in favor of a pro-parceling perspective. In the present article we extend this research, demonstrating why parcel solutions are typically inappropriate when data do not provide a reasonable approximation to the ICM-CFA model.

The present investigation is based on four studies, each of which investigates potential problems with the use of item parcels when the traditional ICM-CFA model does not fit data based on responses to individual items. In each study we show that the use of item parcels can—and typically does—substantially inflate the apparent goodness of fit. These inflated fit indexes suggest that the fit is so good that applied researchers might be led to the erroneous conclusion that misfit was not a problem. In each of the studies, substantive interpretations of interest to pragmatists are biased by the use of parcels.

Studies 1 and 2 are based on real data that address critical issues of substantive importance to applied researchers. Emphasizing the significance of these issues to actual practice, we use data from two of the most widely used instruments in psychology. Study 1 is based on responses derived from the 10-item Rosenberg Self-Esteem instrument collected on four occasions (see Marsh, Scalas & Nagengast, 2010, for more detail). Study 2 is based on responses to two 12-item scales designed to measure Extraversion and Neuroticism (see Marsh, Lüdtke, et al., 2010 for more details)—two of the big-five personality factors. In each case previous research has consistently shown that analyses of item responses

led to unacceptably poor fits that undermined the credibility of these measures, some of the most widely used instruments in psychology. In each case, the use of item parcels has been a popular – but inappropriate – way to deal with these problems. Thus, we show how the use of item parcels merely camouflages these problems in a way that results in an artificially inflated estimate of fit, but does not really resolve the problems. More specifically, we show that more appropriate models (that do not rely on overly restrictive ICM-CFA assumptions or inappropriate use of item parcels) apparently do resolve the problems. In pursuing this aim, both studies offer apparently novel solutions to problems that have plagued applied researchers for decades and have been the basis of dozens of studies in relation to each of these instruments.

Study 3 and especially Study 4 go on to explore implications of these issues with simulated data so that biased parameter estimates can be compared to true population values. Following from Study 2, Study 3 is based on simulated data for two 12-item scales in which the population structure is purely unidimensional (i.e., ICM-CFA), a good simple structure, or a moderate simple structure. Study 4 is a SEM based on simulated data representing a realistically complex structure in which there are three predictor factors predicting a single outcome variable. The structure of the predictor variables is complicated by the existence of a method factor and cross-loadings so that only the outcome factor is purely unidimensional (i.e., ICM-CFA). Simulated data based on a known population generating model are particularly appropriate for testing consequences of using item parcels in combination with ICM-CFA models under varying conditions of misfit. However, unlike many studies that only consider data for which the unidimensionality assumption is met, we systematically contrast results based on purely unidimensional models with models in which this assumption is violated to varying extents.

Study 1: The Factor Structure of Self Esteem

The Rosenberg (1965) Self-Esteem Inventory (RSEI) inventory is one of the most widely used self-report measures in psychology and the social sciences. Yet, there is a substantial and ongoing debate about the appropriate factor structure to represent RSEI responses (e.g., Marsh, Scalas & Nagengast, 2010). Nevertheless, there is wide agreement that responses do not form a pure unidimensional scale. The problem is that positively and negatively worded items do not combine to form a single unidimensional factor. Many studies based on this instrument use scale scores, which simply ignore the problem. A few

early CFA studies (e.g., Marsh et al., 2005) used item parcels in such a way that camouflaged the problem. More recently, researchers have sought to evaluate alternative structures at the item level that dealt explicitly with problems of item wording. As noted by Bandalos (2008), much of this research has focused on how best to model method effects. However, the applied researcher might not even be aware that the method effects exist and thus, may inappropriately use item parcels in a way that inadvertently camouflages the misspecification. Hence, research based on the Rosenberg instrument provides a classic example of inappropriately using item parcels in such a way that a potentially serious problem about the basic structure of the construct in question is camouflaged.

Although there is clear agreement that an ICM-CFA model at the item level is not an appropriate model for Rosenberg self-esteem responses, there is no clear consensus about what alternative model is most appropriate—nor even a clear rationale for how to test competing interpretations. At least three alternative interpretations exist: (a) two substantively distinct self-esteem trait factors (positive and negative self-esteem), (b) one self-esteem trait factor and ephemeral method artifacts associated with positively or negatively worded items, or (c) one self-esteem trait factor and stable response-style method factors associated with item wording. Based on these alternatives, Marsh, Scalas and Nagengast (2010) posited 8 models and tests of these models based on longitudinal data (4 waves of data across 8 years with a large, representative sample of adolescents). Longitudinal models provided no support for the unidimensional model, undermined support for the 2-factor model, and clearly refuted claims that wording effects are ephemeral, but did provide good support for models positing one substantive (self-esteem) factor and two response-style method factors that are stable over time. Their longitudinal methodological approach not only resolved at least some of these long-standing issues in self-esteem research, but also has broad applicability to most self-report surveys with a mix of positively and negatively worded items. Consistent with our empirical pragmatist approach, we also note that their preferred model with one substantive (self-esteem) factor and two method factors corresponds to a bifactor measurement model that Reise (2012) and others argue should routinely be contrasted with the traditional ICM-CFA and related item response theory (IRT) models that assume pure unidimensional factors.

In the present investigation, we extend this research in analyses based on item parcels by considering the nature of conclusions that implicitly assume t no item-wording method factors. This study

is based on Marsh, Scalas, and Nagengast (2010) where the sample, data, analyses, and alternative models are described in more detail (also see Appendix 1, supplemental materials). Briefly, the Youth in Transition (YIT) data (Bachman, 2002) is based on a representative sample of 87 U.S. public high schools and approximately 25 students from each school, collected on four occasions: Wave 1—10th grade (N = 2,213); Wave 2—11th grade (N = 1,886); Wave 3—12th grade (N = 1,799); Wave 4—1 year after normal high school graduation (N = 1,620). For both single-wave and longitudinal CFAs, full information maximum likelihood estimation (FIML) was used to account for missing data (Enders & Bandalos, 2001; Muthén & Muthén, 2010). We focus on only two (see Figure 1) of the eight models considered by Marsh, Scalas, and Nagengast (2010). Model 1 is a pure unidimensional model that posits one self-esteem factor with no method factors associated with item wording. Model 2 also posits one global self-esteem factor, but it includes method factors associated with positively and negatively worded items.

For the purposes of the present investigation, we evaluated the fit of Model 1 based on item parcels as well as responses to the original 10 items. Three parceling strategies were considered. For the 5-parcel data, the first parcel was the average response to the first and sixth of the 10 items, the second parcel was the average response to the second and seventh items, and so forth. For the 3-parcel data, the first parcel was the average response to the first, fourth, seventh, and tenth items; the second parcel was the average response to the second, fifth, and eighth items; the third parcel was the average response to the second, fifth, and eighth items. These 5- and 3-parcel solutions are based on a distributed strategy in which each parcel has a mixture of positively and negatively worded items. However, because of the nature of the items, we also explored a 3-parcel homogeneous strategy in which the four negatively worded items form one parcel, and the six positively worded items form the other two parcels.

Following Marsh, Scalas and Nagengast (2010) we report three goodness of fit indexes routinely provided by statistical packages: Root Mean Square Error of Approximation (RMSEA), the Tucker-Lewis Index (TLI), and the Confirmatory Fit Index (CFI)], as well as the robust $\chi 2$ test statistic and an evaluation of parameter estimates. The TLI and CFI vary along a 0-to-1 continuum, and values greater than .90 and .95 typically reflect acceptable and excellent fits to the data. RMSEA values of less than .05 and .08 reflect a close fit and a reasonable fit to the data, respectively (Marsh, 2007; Marsh, Hau, & Grayson, 2005). However, we emphasize that these cut-off values only constitute rough guidelines (Marsh, Hau & Wen,

2004; also see Marsh, 2007; Marsh, et al., 1998, 2005). Consistently with recommendations for longitudinal panel data more generally (Marsh & Hau, 1996; Jöreskog, 1979), correlated residuals were posited a priori between matching indicators of self-esteem in the longitudinal data (see Marsh & Hau, 1996).

Results and Discussion: Study 1

Analyses of item responses. Model 1 was not able to provide an acceptable fit to the data when applied to item-level data. Next we evaluated Model 2 for item-level data, the preferred model based on the Marsh, Scalas, and Nagengast (2010) study. In relation to traditional interpretations of fit indexes, Model 2 provides an excellent fit to the data for both the analyses of each wave considered separately (see Appendix 2, supplemental materials) and for the longitudinal data that is the focus of Study 1 (Table 1). When Model 2 was applied to longitudinal data, there was substantial stability for self-esteem; test-retest correlations for adjacent waves varied between .71 and .82. However, there was also substantial stability of the positive and negative item factors (test-test correlations for adjacent waves varied between .49 and .60). The stability of these method factors is substantively important, demonstrating conclusively that these method-effects are not short-term ephemeral method artifacts. In Model 2, which was applied to longitudinal data, it was also possible to constrain the correlations between method effects over time to be zero (consistent with ephemeral method effects) or to allow them to be freely estimated (consistent with the stable, response-style explanation). However, constraining these correlations to be zero led to a substantial decline in goodness of fit. Although not a focus of the present investigation, there is good support for the invariance over time of factor loadings and factor variances, and—to a lesser extent—item uniquenesses (see Marsh, Scalas & Nagengast, 2010).

As the goodness of fit for Model 1 is completely unacceptable in relation to current standards, it is dubious even to consider parameter estimates based on this model. Nevertheless, it is interesting to note that estimates of the stability of self-esteem over time for the longitudinal data are somewhat smaller for Model 1 than for Model 2. This occurs because the item wording effects that are confounded with the self-esteem construct in Model 1 are less stable than the self-esteem factor itself. Had the item wording effects been less stable over time, the stability estimates based on Model 1 would have been even lower. However, the most important message from the evaluation of Model 1 is that a single unidimensional factor does not

fit the data and reliance on it would have undermined our ability to evaluate the positive and negative item wording effects—the main focus of the Marsh, Scalas and Nagengast (2010) study and a considerable body of research reviewed in their article.

Analyses of parcel scores. Our main focus here is what happens when analyses for Model 1 are based on parcel scores rather than items? For the 5-parcel data with distributive parceling, the fit of Model 1 is excellent according to current standards and would typically lead to the acceptance of this model. Similarly, there is good support for the invariance over time of the factor loadings, factor variances, and—to a lesser extent—the uniquenesses. For the 3-parcel data with distributive parceling, the goodness of fit is even better—again leading to the conclusion that Model 1 does a very good job of describing the responses. Finally, the 3-parcel solution based on the homogeneous parceling strategy (placing the four negatively worded items in one parcel, and the six positively worded items in the other parcels) resulted in a noticeable decline in fit relative to the other 3-parcel model (based on the distributive strategy). However, by conventional standards the fit of this model was still good, and clearly better than models based on item responses (or even the 5-parcel model). It is also informative to compare the stability coefficients for self-esteem based on this parcel data. The results (Table 2) demonstrate that stability coefficients are very similar for Model 1 based on responses to 10 items, five parcels, and both of the three-parcel solutions. Hence, interpretations based on Model 1 for analyses of items (and its comparison to Model 2) apply as well to the analyses of parcel data.

These results suggest that in this application, the use of parcels did not systematically affect the test-retest correlations, but substantially inflated goodness of fit, camouflaged longitudinally stable method effects evident at the item level, and undermined understanding of the factor structure for this widely used instrument. Because the fit of these parcel solutions was so good, the unwary applied researcher might be led to believe that there was no misspecification in the model. Reliance on parcel solutions would not have allowed us to resolve this substantively important issue about the item wording effects for this widely used instrument. The fact that the test-retest correlations were relatively biased in this example suggests that parcels might not bias conclusions based on relations at the construct level – a focus of the pragmatist. However, as we will show in the other studies, this interpretation should be viewed with caution as Study 1

was based on highly restricted example involving a single substantive construct measured over time, rather than multiple constructs potentially affected by additional sources of misspecification.

Study 2: The Factor Structure of Personality Measure (Neuroticism and Extraversion)

Study 1 was based on a single construct and on misspecification due to method effects associated with items designed to measure this single construct. However, a more typical misspecification problem with the ICM-CFA model occurs when more than one construct is considered. In Study 2 we illustrate these problems with one of the most widely studied factor structures in psychology, the big-five personality factors. This is particularly relevant as much of this research is based on factor analysis, but the use of CFA is seen as controversial or even problematic. Thus, for example, big-five personality researchers (McCrae, Zonderman, Costa, Bond & Paunonen, 1996, p. 563) concluded:

> *In actual analyses of personality data from Borkenau and Ostendorf (1990) to Holden and Fekken (1994), structures that are known to be reliable showed poor fits when evaluated by CFA techniques. We believe this points to serious problems with CFA itself when used to examine personality structure the ICM-CFA model is inappropriate.*

Marsh, Lüdtke, et al. (2010) also argued against the appropriateness of ICM-CFA models, but they demonstrated that ESEM solutions at the item level provided a substantially better fit to the data (incorporating items that load on more than one factor) and resulted in systematically smaller factor correlations. They posited and demonstrated that specifying positive cross-loadings to be zero resulted in a misspecified model in which ICM-CFA factor correlations were systematically inflated in a way that undermined support for discriminant validity and biased estimated relations with other factors (also see Marsh, Muthén, et al., 2009). Marsh, Lüdtke et al. were also critical of the use of item parcels, noting that many applied big-five researchers use ICM-CFA models in combination with item parcels. For example, Sterba and MacCallum (2010) used an ICM-CFA of big-five data in combination with item parcels to provide an empirical demonstration of what they presented as best practice and potentially serious problems in the use of item parcels. Although Marsh, Lüdtke et al. did not pursue this issue, we noted earlier that alternative parceling strategies are likely to result in substantively different conclusions when the unidimensionality assumption underlying the construction of parcels is violated. The juxtaposition of these conflicting claims and widespread practice in big-five research provides a particularly heuristic basis

for tests of these alternative perspectives. Thus, in Study 2 we juxtapose ICM-CFA and ESEM solutions based on responses to items with solutions based on item parcels.

The data for Study 2 are responses to 24 items that measure Neuroticism and Extraversion based on the German version (Borkenau & Ostendorf, 1990) of the NEO-FFI big-five personality instrument (Costa & McCrae, 1992). Data came from the Marsh, Lüdtke, et al. (2010) study, which demonstrated the usefulness of ESEM with a large representative sample of upper-secondary students (and presented details of the sample and a priori model in more detail; also see Supplemental Materials). In particular, they found that ICM-CFA models provided a systematically poorer fit to the data than corresponding ESEMs, and demonstrated logically and empirically that correlations among the personality factors were positively biased in the ICM-CFA models, and differed systematically compared to those based on the corresponding ESEM models. Marsh, Lüdtke, et al. argued that this was a potentially serious problem to the pragmatist in that relations among the personality factors—and their relations with other constructs (e.g., background variables, covariates, or subsequent outcomes)—are the primary focus of most applied personality research. In the present investigation, we extend this study to determine what happens when analyses are based on parcels rather than individual items.

An interesting complication in study 2 is the inclusion of correlated uniquenesses (CUs) that are easily incorporated into CFA and ESEM models, but are not easily incorporated into traditional EFAs like those used with big-five data as an alternative to ICM-CFA models or the construction of item parcels. CUs typically should be avoided, but in some circumstances they should be included a priori (e.g., Jöreskog, 1979; Marsh & Hau, 1996). As described in more detail by McCrae and Costa (2004), in the NEO-PI-R (full, long version with 240 items) each of the big-five factors is represented by six facets and each facet is represented by multiple items. However, for the NEO-FFI, 12 items were selected to best represent each of the big-five factors without reference to the facets, so that some facets are over-represented (relative to the design of the full version of the instrument) whilst other facets are represented by a single item or not represented at all. Marsh, Lüdtke et al. (2010) showed that items from the same facet of a particular big-five factor have higher correlations than items that came from different facets of the same big-five factor—beyond correlations that could be explained in terms of the common big-five factor that they represented. They modeled these potentially inflated correlations due to facets, as CUs. They showed that ICM-CFA

and ESEM solutions that included these CUs fitted the data systematically better, and that the CUs were invariant over gender and time. In the present investigation, all solutions based on items included these CUs, but CUs could not be modeled in analyses of parcels. Although the inclusion of CUs had little effect on factor correlations, the inability to include them in analyses of parcels points to another limitation of parceling. It is also worth noting that the inclusion of CUs differentiated the ESEM solution from a traditional EFA solution that represents a special case of the more general ESEM framework.

Analyses based on Mplus 6 consisted of traditional ICM-CFAs and ESEMs based on the Mplus robust maximum likelihood estimator (MLR) with standard errors and tests of fit that are robust in relation to non-normality of observations (Muthén & Muthén, 2010). The ESEM approach differs from the typical CFA approach in that all factor loadings are estimated, subject to constraints necessary for identification (for further details of the ESEM approach and identification issues, see Asparouhov & Muthén, 2009; Marsh et al., 2009). For present purposes, we chose to apply the target rotation in which items were given a target value of zero on the factor that they were not intended to represent, and the deviation from this factor loading pattern was minimized. As emphasized by Browne (2001; also see Asparouhov & Muthén, 2009; Dolan, Oort, Stoel & Wicherts, 2009) this strategy reflects a compromise between the rationale of EFA and ICM-CFA, based on partial knowledge of the factor structure and also reflects the logic underlying evolving Bayesian models as operationalized, for example, in BSEM procedure in Mplus (Muthén & Asparouhov, 2012). However, the important difference is that in ICM-CFA, factor loadings specified to be zero are forced to assume this value. In contrast, for target rotations the factor loadings specified to be zero are made to be as close to zero as possible, but they are not constrained to be zero. Thus, whilst the targets influence the final rotated solution, the factor loadings for target items can end up as very different from zero if the zero loadings are not appropriate (Asparouhov & Muthén, 2009). Based on simulated data, Asparouhov and Muthén (2009) suggested that target rotation is particularly appropriate when there is a clearly defined a priori factor structure and at least a reasonable approximation to simple structure.

We began by comparing 8 models, 4 ICM-CFAs and 4 ESEMs. The initial models posited two a priori factors (e.g., Extraversion and Neuroticism) for responses to the 24 individual items based on ICM-CFA and ESEM. The remaining models were based on item parcels with 6, 4 or 3 parcels (based on the average of responses to 2, 3, or 4 items per parcel, respectively). Item parcels were constructed arbitrarily by

assigning the first item to the first parcel, the second item to the next parcel, and so forth (see Supplemental Materials, Appendix 3 for the Mplus syntax used to test item and parcel solutions using CFA and ESEM).

As in Study 1, we emphasize that typical cut-off values of what constitutes an acceptable fit are only rough guidelines. Although these same criteria have typically been applied to ESEM studies (e.g., Marsh et al., 2009, 2010), even more caution is needed in their interpretation, as there is still relatively little systematic study of their appropriateness to ESEM. Indeed, the ESEM model is considerably less parsimonious than the ICM-CFA model in relation to the number of factor loadings, so that issues of capitalizing on chance and parsimony corrections are likely to be more important.

Results and Discussion: Study 2

ICM-CFA vs. ESEM. We begin with critical analyses to compare corresponding ESEM and ICM-CFA models in terms of goodness of fit and parameter estimates—particularly the size of factor correlations, which is a major focus of the present investigation and especially relevant to a pragmatist perspective to parceling (Table 3; also see Supplemental Materials, Appendix 4). Whilst the ICM-CFA model is not able to fit the data adequately (e.g., CFI = .866, model RDCFA1 in Table 3) the fit of the corresponding ESEM is better and minimally acceptable (e.g., CFI = .921, model RDESEM1 in Table 1) according to current standards. This difference in the goodness of fit calls into question the appropriateness of the ICM-CFA model for these data. Furthermore, the systematically improved fit for the ESEM model and resulting parameter estimates suggest that the items are not unidimensional; there are systematic cross-loadings of items designed to measure one factor, on the other factor. As described earlier, this pattern of results also calls into question the appropriateness of using parcel scores instead of items.

Substantively, it is important to evaluate the sizes of factor correlations. Of particular relevance, the factor correlation is substantially smaller for the ESEM solution (r = .15) than for the corresponding ICM-CFA solution (r = .51). In this respect, the ESEM solutions –compared with the ICM-CFAs solution– are more consistent with theoretical predictions that the big-five personality factors are reasonably orthogonal. Marsh et al. (2010) also described logically why requiring the small positive cross-loadings associated with big-five items to be zero results in inflated factor correlations. When small positive cross-loadings are constrained to be zero, these non-zero relations between an item and a non-target factor can only be represented through the factor correlations so that estimated factor correlations are positively

biased. Systematically inflated correlations for the ICM-CFA solution have potentially serious implications for applied research where differential validity is important (e.g., distinctive profiles for personality factors or applications when all five factors are related to outcome variables such that issues of multicollinearity become important).

Parcel Indicators. Next, we examine the use of parcels for these data. The 24 items were used to form 12 parcels (of 2 items each), 8 parcels (of 3 items each), or 6 parcels (of 4 items each). Hence, each of the two latent factors was represented by 6, 4, or 3-parcel indicators. The use of parcels, as expected, substantially improved the goodness of fit indicators and the sizes of the factor loadings, but had little effect on the size of the factor correlations. Indeed, the 6-parcel ICM-CFA solution has such good fit indexes (e.g., CFI = .994, TLI = .989, RMSEA = .015), that most applied researchers would readily accept this as an appropriately fitting model. However, results based on the 24 items clearly show that the ICM-CFA model is not appropriate and that the large estimated correlation between the two latent factors is substantially inflated in relation to the ESEM solutions and theoretical predictions. In this respect, the use of parcels camouflaged the misfit in the ICM-CFA, but had no effect on the size of the inflated correlation associated with the failure to take into account cross-loadings at the item level.

One-factor solutions. As emphasized by Bandalos, applied researchers frequently have to decide whether a construct posited to be unidimensional actually reflects two or more underlying factors (or, perhaps, subfacets of a more general factor) and whether different parceling strategies can distort interpretations in relation to this question. This was clearly an important issue in Study 1, but is not a particularly relevant concern for the big-five data considered here as the separation of the two latent factors is well established theoretically and empirically. Indeed, the one-factor solution (Model RD1CFA1A in Table 3) based on 24 items clearly provides a poor fit to the data (e.g., CFI = .636) and is easily rejected. [We also note that ICM-CFA and ESEM solutions are equivalent for one-factor models].

What happens when we evaluate the one-factor solution with different parceling strategies? For purposes of convenience, we first reverse scored all the items in the Neuroticism factor so that the two factors were positively correlated (positively oriented Neuroticism is typically referred to as emotional stability). One strategy would be to simply re-use the parcels used in the analysis of the two-factor solutions. In relation to the one-factor model these represented a homogenous parceling strategy that did

not confound the factors (i.e., the first half of the parcels were associated with the first latent factor and the second half were associated with the second latent factor; these are labeled as "homogeneous" one-factor parcel solutions in Table 1). Each of the (12, 8, and 6) parcel solutions resulted in a poor fit. Thus, like the 24-item solutions, these parcel solutions lead to the rejection of the one-factor hypothesis.

For the purpose of the distributive parceling strategy we used a typical sequential approach to assign items to factors. Thus, for example, for the 12-parcel solution, the 1st and 13th items were assigned to the first parcel and so forth until the 12th and 24th items were assigned to the 12th item parcel. Importantly, this apparently reasonable parceling strategy completely confounds the Extraversion and Neuroticism factors (i.e., each of the 12 item parcels would have one Extraversion and one Neuroticism items). Furthermore, the extent of confounding varies somewhat with the number of item parcels in a way that is idiosyncratic to the particular application. As applied here, the two factors are completely confounded for the 12- and 6-parcel solutions, but only partially confounded in the 8-parcel solution (i.e., the first four parcels would have two items from the first factor and only one from the second, and vice-versa for the last four parcels). Because this source of misfit is completely or partially camouflaged by the construction of parcels, the fit of models based on these the parcels (Models RD1CFA2b- RD1CFA4b in Table 3) are substantially improved. Indeed, all three of these solutions might be argued to be acceptable according to some criteria of acceptable fit, and applied researchers might claim (inappropriately) that the responses to the 24 items—as reflected in the parcel scores—can be represented by a single latent factor.

Summary and implications. Big-five personality factors have dominated recent personality research but the failure of ICM-CFA models to provide acceptable fits to big-five responses has been a serious limitation of this research. Marsh, Lüdtke, et al. (2010) demonstrated that the problem—at least in part—is reliance on the ICM-CFA model, which requires items to load on one and only one factor (i.e., to be purely unidimensional) and showed that ESEM apparently resolves this problem. Of particular substantive importance to the applied researcher, failure to account for cross-loadings in the ICM-CFA model leads to systematic biases in the estimated factor correlations. Here we extend this analysis to explore the implications of different parceling strategies that fail to take into account cross-loadings, such as those that were evident in the item-level analyses. For all two-factor solutions, the parcel strategies greatly improved apparent goodness of fit and in many instances led to such good estimates of fit that the

the unwary applied researcher might be led to believe that cross-loading misspecification was not a problem. However, each of these parcel solutions resulted in approximately the same inflated estimate of the factor correlation as the misspecified model based on analyses of items. In this sense the use of parcels merely camouflaged the misspecification (in relation to goodness of fit and empirical parameter estimates), and did not resolve it. Worse, it could lead unwary researcher to the conclusion that these two factors are highly correlated, a correlation sufficiently high to bias estimates between big-five factors and other outcomes. In some cases, the correlations may be even so high as to detract from the discriminant validity of the factors, leading the researcher to explore potential one-factor solutions.

Indeed, although one-factor solutions are clearly inappropriate for these data, we used these models to demonstrate why homogeneous and distributive parceling strategies differ so fundamentally when the assumption of unidimensionality is violated. So long as we used a homogeneous parceling strategy (in which no parcels contained both extraversion and neuroticism items), the one-factor model was clearly rejected in relation to its poor fit to the data. However, when a distributive parceling strategy was used such that extraversion and neuroticism items were completely confounded by the parcels, there was reasonable support for a one-factor model. Although it might be argued – and we would agree – that it would be illogical to form parcels with a mix of items from different factors, this is precisely the logic of the distributive approach to parceling. Although clearly inappropriate in the present application, the logic of the distributive approach to parceling is also suspect whenever there is a systematic source of misfit in the ICM-CFA model based on item responses.

From a pragmatist perspective, the estimated correlation between these two big-five factors varied widely, depending on the model: .15 for the ESEM model with cross-loadings; .49 for the ICM-CFA model that did not allow cross-loadings; between .47 and .52 for the two-factor parcel solutions that ignored the cross-loading misspecification; and an implicit 1.0 for the one-factor solution based on the distributive parceling strategy. Importantly, with the exception of the ICM-CFA model based on item responses, all these models resulted in goodness of fit values that met traditional standards of acceptable fit to the data. In summary, Study 2 demonstrates that the use of parcels can camouflage misspecification in such a way that substantially biases estimates of relations between constructs and should be seen as unacceptable even from a pragmatist perspective.

Study 3: CFA, ESEM, and the Use of Item Parcels with Simulated Data

The use of real data is important in terms of demonstrating the practical significance of our concerns for actual applied research. However, there are also important advantages to the use of simulated data for which the true population generating model is known. In Study 3 we simulated data for 24 items based on factor structures that can be described as what Sass and Schmitt (2010) refer to as approximate simple structure (see supplemental Appendices 5 and 6 for a summary of the population generating models). Simulated data were generated using the true factor pattern loadings for item *i* on Factor 1 or 2 (λ_{i1} & λ_{i2}), the factor correlation (ρ), and the item residual (ε_i). In order to simplify interpretations, items were standard normal variables with residual variances defined as $\text{Var}(\varepsilon_i) = 1–[(\lambda^2_{i1} + \lambda^2_{i2} + 2\ (\lambda_{i1} \times \lambda_{i2} \times \rho)]$. For each of the factor structures, the population generating model had a factor correlation of ρ =.25 or .60. The factor correlation of .25 was chosen to be moderate and representative of many applied applications. The factor correlation of ρ =.60 was chosen to be sufficiently large that it might be (inappropriately) considered reasonable to posit a single underlying factor. Indeed, it is not uncommon in applied psychological research to have distinct constructs that are correlated as high as .60—sometimes based on factors that are supposed to be distinct and sometimes based on what is supposed to be a single underlying factor.

Four population generating models were considered (see Supplemental Materials, Appendix 5). The first was a pure ICM-CFA model in which each of the 24 items loaded on one and only one factor. The second was nearly a pure ICM-CFA model in that all cross-loadings varied between 0 and .10. We consider this a closer approximation to pure unidimensionality than applied researchers are likely to encounter in actual practice. In the third model, a good approximation to unidimensionality, items had small cross-loadings (four cross-loadings each were 0, .10, and .20). However, this is a very good approximation to simple structure, and better than might be expected in many applied studies. The fourth was a moderate approximation to unidimensionality (two cross-loadings were 0, one was .10, and three each were .20, .30, and .40) such that all cross-loadings were smaller than the corresponding loading on the factor it was designed to measure.

In summary, there were 8 simulated sets of data: four population generating structures (pure, near perfect, good, and moderate) ×2 factor correlations (.25 and .60). Each simulated dataset contained a single

sample of 100,000 (see footnote 1) simulated cases based on a known population-generating model (see supplemental materials). The intent of analyses based on these datasets is to determine how well the ICM-CFAs and ESEMs based on items and parcels could fit data from this population generating model, rather than evaluating how this picture was clouded by sampling variation per se (but see Sterba, 2011; Sterba & MacCallum, 2011). For this reason we evaluate the structure based on one large population-like sample rather than evaluating the performance of the model based on many samples that vary in terms of sample size. ICM-CFA and ESEM models applied to these data were similar to those applied to the big-five data. Again, all analyses were conducted with Mplus 6 (Muthén & Muthén, 2010), consisting of traditional ICM-CFAs and ESEMs with target rotations based on a robust maximum likelihood estimator (MLR).

Results and Discussion: Study 3

For the simulated data in Study 3, we began with known populations generating models. When the ICM-CFA model was used to generate the data (the pure unidimensional structure), solutions based on the 24 items and each of the parceling strategies all were able to fit the data and accurately estimate the known factor correlation. Hence, the use of parcels is appropriate when data are purely unidimensional.

However, we note that real data in applied research never are purely unidimensional, so we now turn to a nearly perfect unidimensional structure that is still closer to a unidimensional ideal than applied researchers are likely to find in practice. For this data, the goodness of fit for the ICM-CFA model based on 24 items is extremely good (CFI=.990, TLI=.989, RMSEA=.016; see Table 4) and would typically lead the applied researcher to readily accept this model as providing a good fit to the data. Nevertheless, due to the failure to take into account the (very small) cross-loadings, the estimated factor correlations are clearly inflated in relation to the known population correlation: r = .41 (for ρ =.25) and r = .71 (for ρ =.60).

We now consider the good and moderate structures that are more representative of structures likely to be found in applied research. Even here, all the misspecified ICM-CFA solutions based on 24 items provided an apparently good fit to the data (all TLIs and CFI > .94, RMSEAs < .03; see Table 4). However, in these misspecified models that do not consider cross-loadings, the estimated factor correlations are substantially inflated in relation to the known population correlation: rs = .52 (for ρ =.25) and r = .78 (for ρ =.60) in the good structure; rs=.84 (for ρ =.25) and .94 (for ρ =.60) in the moderate

structure. This result is important since none of the population-generating models that we used included cross-loadings so large as to be considered highly atypical in applied practice.

For each of the eight simulated datasets, the ESEM solutions provided an almost perfect fit to the data (e.g., all TLIs and CFI ~ 1.0, RMSEAs ~0; see Table 4). For the pure unidimensional datasets, the goodness of fit measures for the ESEM and ICM-CFA solutions were comparable and both accurately estimated the population factor correlation. However, the ICM-CFA was preferable on the basis of parsimony (i.e, the df was higher) and because the ESEM model had many superfluous parameter estimates that were zero in the population generating model. However, the ESEM solutions were preferable for even the model with a near-perfect simple structure, and particularly for the moderate and good structures, even though the fits of the ICM-CFA solutions were apparently acceptable or very good for all these models. The superiority of the ESEM solution is highlighted by the differences in the estimated factor correlations. In particular, the ESEM solutions provided approximately accurate estimates of the true population correlation in all four datasets, whilst the ICM-CFA estimates were substantially biased for the nearly perfect and particularly for the good and moderate structures.

Parcels. Next we examined the use of parcels for these data, using the same approach to parceling as in Study 2. Even though the ICM-CFA solution for 24 items provided an apparently good fit to the data for the good and moderate structures, the fit was systematically improved by the use of parcel scores (Table 4). However, the factor correlations shown to be substantially biased for the ICM-CFA solutions were also substantially biased to approximately the same extent for the parcel solutions.

One vs. Two factor solutions. When the population correlation was $\rho = .25$, the ICM-CFA solution based on 24 items led to rejection of the one-factor hypothesis for all four factor structures (see Table 4). Similarly, the one-factor model was unable to provide a reasonable fit to the data for any of the homogenous parcel solutions in which the parcels were unconfounded with the latent constructs (i.e., parcels were constructed from items within each of the factors so that no parcels had items from different factors). However, for distributive parcel solutions that partially or completely confounded the factors (i.e., parcels were constructed from items from different factors) the one-factor models provided excellent fits to the data. Indeed for the 12 and 6 parcel solutions (where the confounding was complete), the goodness of

fit was almost perfect. In contrast, for the 8-parcel solutions, where the confounding was partial, the fit of the one-factor model was not perfect, but still satisfactory (CFIs and TLI > .96) by current standards.

When the population correlation was ρ = .60, the fits of the one-factor models based on 24 items were not acceptable for the pure, near perfect, and good structures. However, for the moderate structure the fit of this model was reasonable (e.g., CFI=.961, TLI = .953). Solutions based on the homogeneous (unconfounded) parcels for the moderate structure resulted in marginally poorer fits to the data, but the fit statistics were still acceptable according to some commonly-used criteria of an acceptable fit.

For the distributed (confounded) parcels, the 12- and 6-parcel (completely confounded) solutions resulted in almost perfect goodness of fit statistics. Although the fit of the partially confounded (with 8 parcels in Table 4) solution was not quite as good (CFI=.999, TLI=.998), these fit indexes typically would be interpreted as highly acceptable.

Summary and implications: Study 3

The results of Study 3 largely confirmed the results of Study 2, but had additional features that contributed to understanding problems associated with the use of parcels. In particular, because data were based on known population generating models, it was possible to demonstrate empirically that there was a substantial positive bias in the ICM-CFA factor solutions based on items and each of the parceling strategies. We argued logically that this was the case in Study 2 as well, but the findings in Study 3 are more definitive. In both Studies 2 and 3 the factor correlations based on the ESEM solutions were substantially smaller than the ICM-CFA solutions, but for the conditions in Study 3 the factor correlations based on the ESEM target rotation closely approximated the correlation in the population generating model.

Unlike Study 2, in Study 3 we constructed simulated data that perfectly met the unidimensionality assumption underlying the application of the ICM-CFA model and the use of parcels. Consistent with the design of these data, both the ICM-CFA and ESEM solutions resulted in good fits to the data and accurate estimates of the parameter estimates—including the factor correlation. Consistent with the unidimensionality requirement underlying the appropriate use of item parcels, all of the item parcel solutions were also able to fit the data and resulted in accurate estimates of the factor correlation. These

results are important, demonstrating that the item parcels can be used appropriately when the assumption of unidimensionality is met in the data.

However, even for the near perfect approximation to unidimensionality in which cross-loadings were apparently smaller than is likely to be found in applied research, the ICM-CFA solutions based on items resulted in biased estimates of the known population correlations. Not surprisingly, the biases were substantially larger in the good and moderate conditions that were designed to be more like structures actually encountered in typical applied research. Furthermore, the goodness of fits of these misspecified solutions based on items were sufficiently good that they might be interpreted as acceptable in applied research. The corresponding solutions based on item parcels were even better in terms of apparent goodness of fit, but the substantial bias in the factor correlations was nearly unaffected. Because the fits for the ICM-CFA factor structures were reasonable, the comparison with the substantially better fits for the corresponding ESEM solutions was important. In line with suggestions that the comparison of ESEM and ICM-CFA solutions provides guidelines for evaluating the unidimensionality of the responses, these comparisons based on both the goodness of fit and on parameter estimates—particularly the factor correlation—are informative. Without reference to the ESEM solution, the applied researcher might well conclude that the data were sufficiently close to being unidimensional that the use of parcels was justified. These results demonstrate that a reasonable fit for the ICM-CFA model is not sufficient to demonstrate that the factor correlations will be adequately estimated and suggests that it should typically be contrasted with the corresponding ESEM solution.

Study 4: CFA, ESEM, and the Use of Item Parcels in a Complex SEM

In Study 1 we evaluated the implications of the use of parcels for camouflaging method effects. In Studies 2 and 3, we evaluated the usefulness of ESEM in demonstrating problems with the use of item parcels when there were cross-loadings. However, the focus of studies 2 and 3 was on the inflation of factor correlations associated with ICM-CFA models when the assumption of unidimensionality is violated, rather than the consequences of this for prediction. Although two of these studies were based on real data, all were comparatively simple applications relative to the complexity of much applied research. In Study 4 we combine and expand upon these themes in a simulation study involving three predictor factors (PF1–PF3) and one outcome factor (see Supplemental Materials, Appendix 6) and evaluate

consequences of item-parceling for prediction. PF1–PF3 are complicated by a combination of method effects associated with PF1 and PF2 (a structure like that considered in Study 1), and small cross-loadings associated with all three predictor factors. Only the outcome factor has a pure unidimensional structure. PF1 and PF2 are substantially correlated with each other ($\rho_{12} = .6$) and each of these factors is moderately correlated with PF3 ($\rho_{13} = \rho_{23} = .3$).

On the basis of this same factor structure, three different population data sets are generated in which the path coefficients relating the latent PFs to the latent outcome factor were varied. In all three data sets, the three PFs explain approximately 50% of the variance in the outcome factor. The path from PF3 was always $\beta_3 = 0.3$, but the paths from PF1 and PF2 differed across conditions ($\beta_1 = 0.306$ & $\beta_2 = 0.306$ in population 1; $\beta_1 = 0.555$ & $\beta_2 = 0$ in Population 2; and $\beta_1 = 0.670$ & $\beta_2 = –0.200$ in Population 3).

The complex relation between PF1 and PF2 is the critical feature in these simulated data. As in Study 3, the correlation between the first two factors was sufficiently high that applied researchers might posit a single latent factor. Furthermore, failure to account for shared method effects and cross-loadings in these two factors would further inflate this correlation. Especially given the suppression effect evident in Population 3 (PF1 and PF2 are positively correlated but have positive and negative effects respectively on the outcome variable), misspecification of the measurement model is likely to substantially bias estimates of the path coefficients as well as relations with other factors. As in Study 3, simulated data (N = 100,000 cases for each population) were generated and all analyses were conducted with Mplus (version 6). Models considered for each population varied primarily in relation to PF1 and PF2. Different models either:

1. included or excluded the method effects in PF1 and PF2 (“Mth” or “No Mth” in Table 5);
2. represented PF1 and PF2 as two separate factors (“two factor” models in Table 5) or combined them to form a single factor (“one factor” models); and.
3. Included cross-loadings (ESEM factors) or ignored cross-loadings (ICM-CFA factors).

For both the one-factor and two-factor ICM-CFA models, two sets of parcel scores were considered—based on distributive and homogeneous parceling strategies.

Results: Study 4

For each of the three populations, the ESEM model positing PF1 and PF2 as separate factors (“Two Factor ESEM” models in Table 5) fitted the data (CFI=TLI=1.0) and accurately estimated the factor

structure for each of the three population data sets. Exclusion of the method factors had relatively little effect on either of the goodness of fit statistics (e.g., CFIs & TLIs > .99 for all three populations) or parameter estimates. Hence, the main focus of the results is on the extent of bias introduced in alternative ICM-CFA models in each of the populations.

Population 1 ($\beta_1 = \beta_2 = 0.306$). In population 1, the ICM-CFA model with method effects provided an apparently excellent fit to the data (CFI, TLI > .98; Table 5). However, this model substantially over-estimated the relation between the PF1 and PF2 (r_{12} =.897 compared to ρ_{12} = .6), leading to the underestimation of each of the three path coefficients (Table 5).

Given the very high estimated correlation between PF1 and PF2, it might seem reasonable to combine PF1 and PF2 into a single factor ("One-factor CFA" models in Table 5). Although this resulted in a noticeable decline in fit, the fit indices were still reasonable by some standards (CFIs and TLIs > .94). The b_1 path of PF1 to the outcome variable was substantially inflated—reflecting a combined effect of PF1 and PF2—but the effect of PF3 (b_3 in Table 5) was not much affected.

How do the various parceling strategies fare with these data? The patterns of bias observed in each of the ICM-CFA models were similar in the corresponding parcel models. However, the apparent goodness-of-fit for the parceling models did differ. Compared to the item solutions, distributive parcel models resulted in a slightly better goodness-of-fit, whilst the fit of the homogeneous parcel model was slightly poorer. Indeed, the fit of the one-factor model based on the distributive strategy had an exceptionally good fit (CFI = .986, TLI = .982), inappropriately suggesting that it was an appropriate model.

Population 2 ($\beta_1 = 0.555, \beta_2 = 0$). In Population 2 the ICM-CFA model with method effects provided an apparently excellent fit to the data (CFI, TLI > .980). However, again this model substantially over-estimated the relation between the PF1 and PF2 (r_{12} = .898 compared to ρ_{12} = .6), leading to the underestimation of each of the three path coefficients (Table 5). However, because of the nature of the population path coefficients in this population, the estimated path coefficients were much more distorted by the misspecification: path b_1 was substantially inflated (b_1 = 0.787 vs. $\boldsymbol{\beta_1 = 0}$.555), path b_2 was substantially underestimated (b_1 = –0.307 vs. $\boldsymbol{\beta_2 = 0}$), and path b_3 was slightly underestimated (b_3 = 0.282 vs. $\boldsymbol{\beta_3 = 0}$.300).

Again, the high value of r_{12} = .896 or .904 might suggest combining these two factors (i.e., r_{12} = 1). Although the fit of the one-factor model was reasonable (CFI = .951, TLI = .947), the parameter estimates were much different from those for the corresponding two-factor model. Thus, both b1 and b3 were underestimated and the variance explained in the outcome variable was noticeably smaller.

Parcel solutions again did not substantially alter the pattern of bias observed in the ICM-CFA results, but the apparent goodness-of-fit for the parcel models was slightly improved with the distributive parcel models and slightly worse for the homogeneous parcel models. Again, the one-factor model based on the distributive strategy had an exceptionally good fit (CFI = .986, TLI = .982), suggesting (inappropriately) that it might be an appropriate model.

Population 3 (β_1 = 0.670, β_2= –0.200). Population 3 represents a classic example of a suppression effect in which the two variables that are positively correlated with each other have opposite effects on the outcome variable. Particularly here, it is critical that the measurement model be appropriate. The ICM-CFA model with method effects provided an apparently excellent fit to the data (CFI, TLI > .980). However, this model again substantially over-estimated the relation between the PF1 and PF2 (r_{12} =.896 compared to ρ_{12} = .6), leading to extremely biased estimates: particularly path b_1 (b_1 = 1.049 vs. $\boldsymbol{\beta_1}$ = **0**.670) and path b_2 (b_2 = –0.647 vs. $\boldsymbol{\beta_2}$ = –0.200). The standardized path coefficient greater than 1.0, coupled with the very high correlation between PF1 and PF2 might reasonably lead the applied researcher (inappropriately) to consider combining PF1 and PF2. However, in the corresponding one-factor model, the variance explained dropped substantially (R^2 = .348 vs. .503) and the differential effects of β_1 and β_2 were lost.

Parcel solutions again did not substantially alter the pattern of results, but the apparent goodness-of-fit for the parceling models was slightly improved with the distributive parcel models and slightly worse for the homogeneous parcel models. Again the one-factor model based on the distributive strategy had an exceptionally good fit (CFI = .985, TLI = .982), inappropriately suggesting that it might be an appropriate model.

Discussion: Study 4.

Implicit in our presentation of the results was the assumption that the applied researcher knew at least the appropriate factor structure a priori—that there were three predictor factors and a method factor

associated with some items from FP1 and FP2. In an area that is well researched, this assumption is probably reasonable. However, even with this knowledge, misspecification due to reliance on the ICM-CFA models led to extremely biased parameter estimates—particularly in Populations 2 and 3. Furthermore, because the fit of the ICM-CFA models appeared to be reasonable, the applied researcher might have simply accepted (inappropriately) these models as good-fitting representations of the data. In this case, our recommendation that ICM-CFA models should be compared to the corresponding ESEM models was important. The noticeably improved fit of the ESEM model compared to the corresponding ICM-CFA model—coupled with the substantially different parameter estimates—provided a clear indication that something was wrong with the ICM-CFA model.

What would happen if the applied researcher did not know the appropriate factor structure? Assume, for example, that the applied researcher posited that PF1 and PF2 reflected a single factor instead of two factors, or was led to believe so from the inflated correlation between them based on ICM-CFA models. For each of the three population datasets, a one-factor model provided a goodness of fit that was at least marginally acceptable, and parameter estimates that might seem reasonable. Furthermore, the construction of parcels based on a distributed strategy resulted in even better fit indices that would typically be interpreted as strong support for the model.

Given imperfect knowledge about the nature of the population factor structure, how could the applied researcher resolve this problem? McDonald (2010), and many others, warned that over-reliance on global indices of fit will often lead to problems. Instead (or at least in addition), he stressed that researchers need to evaluate the residual variances and covariances at the level of the individual item as well as associated modification indices. In Study 4, this strategy would have made it obvious that combining PF1 and PF2 was inappropriate. In particular, in the one-factor ICM-CFA model, all correlations among residuals for the 10 indicators of PF1 were positive, as were correlations among residuals for the 10 indicators of PF2. However, all the correlations among residuals relating PF1 indicators to PF2 indicators were negative. Even if this evaluation did not alert the applied researcher to the precise nature of the misspecification (i.e., which items were indicators of PF1 and PF2), it would identify the nature of the problem. Then the applied research could have explored an ESEM solution in which PF1 and PF2 would appropriately be identified as separate factors.

In Study 4 the use of parcels was clearly inappropriate. It is not so much that they fundamentally altered the solution from that which was obtained based on the ICM-CFA models at the level of individual items. Rather, they made it more difficult to identify misspecification based on the ICM-CFA model. For example, in the case where the researcher thought that the PF1 and PF2 represented a single factor, an evaluation of residuals would readily reveal the inappropriateness of this assumption. However, if this underlying pattern of residuals had been confounded by the use of parcels, this critical information would not be available to alert the applied researcher to the problem.

It is also relevant to note that there was a noticeable difference between the fit of parcel models based on the distributive and homogeneous strategies. As highlighted earlier, this is another warning that there is a violation of the unidimensionality assumption of the ICM-CFA model and thus, that the use of parcels is inappropriate. To the extent that the factor structure is purely unidimensional (as in Study 3), different parceling strategies should all result in approximately the same goodness of fit, although there can still be differences due to sampling variation – especially when samples sizes are small (Sterba, 2011; MacCallum & Sterba, 2010). In Study 4 we took advantage of the known factor structure to implement the homogeneous parceling strategy. However, even without this a priori information, it is possible to apply the homogeneous parceling strategy on the basis of empirical approaches based on correlations among items or, even better, among residuals. In this sense the appropriate construction of homogeneous parcels is similar to the evaluation of residuals. However, if it is possible to implement effectively the homogeneous approach to parceling such that results based on these parcels differ from those in the distributive strategy, then the use of parcels is probably inappropriate in that this is an indication that the assumption of unidimensionality has been violated.

In summary, the results of Study 4 indicate that the ICM-CFA model can seriously bias results of SEMs when the assumption of unidimensionality is violated. In Study 4 the violations of the assumption of unidimensionality were not extreme. In fact, all the ICM-CFA models in each of the populations that posited PF1 and PF2 as separate factors had very good fit indices (CFIs, TLIs > .98) that would typically lead to the conclusion that the ICM-CFA model was appropriate. Indeed, even the ICM-CFA models positing PF1 and PF2 to be a single factor had at least marginally acceptable levels of fit (CFIs, TLIs > .94). When at least the basic structure of the population generating model is known, comparison of the

ESEM and ICM-CFA models readily identified the problem. However, even when PF1 and PF2 were inappropriately combined to form a single factor, an examination of residual correlations and the comparison of results based on the distributive and homogeneous parceling strategies readily identified that this structure was inappropriate, and led to the appropriate factor structure. Although we recognize that such patterns are likely to be more difficult to identify with real data, these approaches to the evaluation of the factor structure are still appropriate.

Summary and Directions for Further Research

The application of the ICM-CFA model in applied psychological research is typically inappropriate in that real data almost never fit the highly restrictive ICM-CFA assumptions. Indeed, it is widely accepted that all statistical models only reflect approximations to reality and are always wrong (e.g., Cudeck & Henly, 1991; MacCallum, 2003; McDonald, 2010; Marsh, Balla & McDonald, 1998; Thissen, 2001; and many others). It thus follows that the assumption of unidimensionality underlying the ICM-CFA model is always wrong when applied to real data. Under these circumstances, the use of item parcels is typically dubious. Thus, the fundamental premise of the present investigation is that in applications when both the assumptions underlying the ICM-CFA are wrong and the use of item parcels is inappropriate, combining these two wrongs does not make a right.

In general, the hypothesis that any one cross-loading is exactly zero is always false, and will be shown to be false from a statistical perspective with a sufficiently large *N*. It thus follows that constraining a potentially large number of cross-loadings to be zero will always lead to misfits. The question is whether the misspecified model is sufficiently good to warrant consideration. However, to make this decision, analyses must be carried out at the item level and not at the level of parcels that camouflage these issues by changing the data. The juxtaposition of ESEM and ICM-CFA models at the item level facilitates this comparison.

More specifically, the results of the present investigation demonstrate, based on real and simulated data, that misspecification due to inappropriate use of ICM-CFA models can be largely or completely camouflaged by the inappropriate use of item parcels. Importantly, this doubly inappropriate practice resulted in substantially biased estimates of latent factor correlations, undermining support for discriminant validity and for usefulness in applied research. Indeed, expanding on a similar point made by Bandalos

(2008), it is possible to argue (inappropriately) that a single factor solution provides a good fit to the data under apparently benign parceling strategies for real and simulated data in Studies 2-4. However, the inappropriateness of this conclusion was only evident when ESEM and ICM-CFA analyses of item responses were compared.

Of course, the use of parcels is not always wrong: For data simulated from an ICM-CFA model, results from all the parcel models were accurate in terms of goodness of fit and in capturing the true population correlation. Nevertheless, here we point to clear examples to demonstrate that two wrongs can seem to make a right when these strategies are misapplied. The more worrisome concern is how common these examples are. Although clearly this is beyond the scope of the present investigation, we suspect that this doubly inappropriate behavior—the use of parcels with inadequate ICM-CFA models—might be widespread across many disciplines of applied psychology, and even more so when a single parcel is used (i.e. aggregated scale scores) to represent a construct.

ESEM Solutions

The results of the present investigation add to growing support for the use of ESEM instead of—or in addition to—the traditional ICM-CFA model (Asparouhov & Muthén, 2009; Marsh et al., 2009, 2010, 2011a, 2011b). Although it is not yet used widely used, support for ESEM is already established by a number of empirical demonstrations based on real data. Consistent with the present investigation, this previous research has typically found that ESEM solutions provide a better fit to the data than corresponding ICM-CFA solutions. Importantly, a consistent finding has been that ESEM results in systematically lower factor correlation estimates—even when the fit of the corresponding ICM-CFA model is seemingly good. Furthermore, even when ESEM was fitted to data generated by a pure ICM-CFA model, the ESEM goodness of fit and parameter estimates were unbiased (although the ICM-CFA model would be preferred on the basis of parsimony). From this perspective, we suggest that it is useful to routinely apply both ESEM and ICM-CFA models to the same data. To the extent that the ESEM solution fits the data significantly better and there are substantively meaningful differences in the results, the unidimensionality assumption of the ICM-CFA solution is likely to be violated and the use of item parcels is likely to be inappropriate.

There is, however, a potentially important limitation with ESEM solutions, in that the pattern of cross-loadings and size of the estimated factor correlation will vary with the specific details of the rotation (e.g., Browne, 2001; Sass & Schmitt, 2010). Because of the inherent rotational indeterminacy in EFA (Mulaik, 1972) and ESEM, models based on different rotations are all able to fit the data equally well, so that goodness of fit provides no basis for choosing the most appropriate rotation (Sass & Schmitt, 2010; Schmitt & Sass, 2011). In the present investigation with simulated data, the target rotation provided accurate estimates of the factor correlations. However, this assessment was based on knowledge of population generating models in which some of the items representing each factor had zero cross-loadings in the population generating model and were the target items in the target rotation. Although it is common in factor analysis for several of the indicators to serve as "markers" of the factor (e.g., Cattell, 1949; Comrey, 1984; Gallucci & Perugini, 2007; Howarth, 1972; Overall, 1974), this is clearly not the case in all situations, and the target rotation would not be expected to perform as well—at least in terms of accurately estimating the true population correlation. However, we also note that this assumption is far less extreme than the ICM-CFA model that assumes that all items have zero factor loadings on all factors other than the one they are designed to measure (i.e., that the unidimensionality assumption is met).

Ironically, the historical rationale for most rotation strategies has been on maximizing the simple structure of the factor loadings (for either variables, factors, or a combination of the two) with little regard to the appropriateness of factor correlation estimates (Sass & Schmitt, 2010). Indeed, as emphasized by Sass and Schmitt, there is necessarily a balance between constraints on the sizes of cross-loadings and factor correlations (the extreme being the ICM-CFA solution that constrains cross-loadings to be zero and typically results in substantially inflated factor correlations when this assumption is violated). Although resolution of this problem of choosing the most appropriate rotation strategy is clearly beyond the scope of the present investigation, it is important to emphasize that the goodness of fit for the ESEM solution does not depend on the rotation. Hence, if the fit of the ESEM solution is substantially better than that of the ICM-CFA solution—no matter what rotation is used, the estimated correlation for the ICM-CFA solution is likely to be substantially biased and the use of item parcels with an ICM-CFA model is likely to be inappropriate.

The Use of Item Parcels

In the present investigation, we have focused on situations in which the use of item parcels is most justified—single-group, single-level analyses based on well-established factor structures that provide at least a reasonable approximation to simple structure. As emphasized by many others (e.g., Bandalos & Finney, 2002; Little, et al., 2002; Marsh, et al., 1998), the use of item parcels is even less justified for scale development studies or studies not based on well-established instruments. We also note that the use of item parcels is logically inappropriate for multiple group studies seeking to evaluate differential item functioning and levels of measurement invariance needed to justify the comparison of group means (latent or manifest). Similar arguments exist for tests of invariance over time and latent means in longitudinal data when the same measures are administered on multiple occasions. A critical assumption underlying these mean comparisons is the absence of differential item functioning (Meredith, 1993; Meredith & Teresi, 2006; Marsh et al., 2009, 2010)—that differences in factor means are represented in each of the items associated with the factor. Clearly, the use of item parcels is inappropriate in studies seeking to evaluate differential item functioning over time or across groups, in that violations at the item level are likely to be camouflaged by item parcels (Marsh et al., 2010). In related work, Meade and Kroustalis (2006) reported that for simulated data with a known lack of invariance, models using parcels as indicators more often erroneously indicated that measurement invariance existed than models using items as indicators. We also note the empirical approaches to creating homogeneous parcels in relation to factor loadings would not necessarily result in homogeneous parcels in relation to item intercepts or item uniquenesses. Hence complications in the construction and interpretation of item parcels are likely to multiply as researchers apply them to more statistically complex and substantively interesting models.

What are the proposed advantages for the use of item parcels and support for these claims?

1. Clearly, item parcels are more reliable, result in larger factor loadings, and have smaller residual variance estimates than the corresponding items. These results are clearly evident in the present investigation. However, these differences are largely illusory since we remove from the model of individual scores based on multiple indicators and replace these by a single parcel score. Hence scale or factor reliability based on items and parcels is likely to be similar. From the perspective of generalizability and estimation precision, it is good to have more items – whether the latent factors or aggregated scale scores are based on items or item-parcels. Indeed, because item parcels are based on an implicit assumption

that all items within the parcel are weighted equally, factors based on item parcels might be less reliable if this assumption is violated. In this sense, the use of item parcels is most justified when all factors are purely unidimensional, and when all indicators of each factor are at least essentially tau-equivalent (i.e., equal loadings) or perhaps even essentially parallel (equal factor loadings and uniqueness) or parallel (equal loadings, uniquenesses, and intercepts). Hence, as tests of measures become more sophisticated in relation to problems of differential item functioning, the use of item parcels becomes increasingly problematic. Support for unidimensionality in relation to factor loadings does not justify the use of item parcels in studies of differential item functioning, latent means, and tests of measurement invariance, but these issues have not been widely considered in the parceling literature.

2. Item parcels are claimed to result in fewer violations of normality assumptions. Although not a focus of the present investigation, Hau and Marsh (2004) evaluated the use of item parcels as a strategy for overcoming nonnormality problems with simulated data with extreme skew and kurtosis, but found only modest support for the use of parcels to counter this problem. Similarly, item parcels are sometimes used for Likert responses based on a limited number of response categories. However, more appropriate estimation methods for non-normal data and ordered-categorical estimators are now readily available in most commonly used statistical packages and do not require substantially larger *N*s than maximum likelihood estimation (see discussion by Bandalos, 2008). Furthermore, it is also possible to test ESEMs based on categorical data and to compare these with corresponding ICM-CFA solutions.

To further illustrate how parcels may be problematic in such cases, consider an example with 7 dichotomous items: that is, there are 128 different ways (2^7) to respond to these 7 questions. If an item parcel is formed, these 128 different response patterns are reduced to 8 *(*0,...,7*)* distinguishable responses, no matter whether raw scores, average scores, or something else is used to "parcel" the responses. This is a reduction of complexity by a factor of 16, which is based on the assumption that essentially all response patterns that lead to the same raw score represent the same amount of information about the latent variables involved. If multiple parcels are used, the reduction factor is multiplied for as many parcels as there are in the reduced data. This reduction obviously leads to an impaired ability to detect interactions between latent variables that affect some but not all items or cross-loadings on individual items (since they do not exist anymore in the reduced—parcel level—data). From the perspective of categorical data analysis, the

transformation from item-level data to parcels comes with a substantial change in the multidimensional table that is being analyzed, and comparisons between models for these very different tables would amount to the proverbial comparisons between apples and oranges.

A somewhat related issue is a potential misconception of proponents of classical test theory (CTT): Instead of using item-level models for categorical data, CTT essentially produces one single item parcel by computing the sum score over all items in a test. Then, proponents of this method argue, CTT can be used even if item-level models show violations of unidimensionality or other types of misfit when applying for example, item response theory (IRT) or log-linear models with latent variables. This conclusion is obviously a fallacy, since the data reduction performed by producing the CTT test score goes from a *k*-dimensional table with m^k cells for *k* items to a unidimensional table with $mk+1$ $(0,\ldots,mk)$ cells. For example, if 10 items with 4 categories are analyzed, the 10-dimensional table has $4^{10} = 1048576$ cells, which are reduced to 41 (0,…40) distinct raw scores (see von Davier, 2010). A data reduction of this magnitude leads to a unidimensional table which—one could say by definition—does not allow the detection of violations of unidimensionality. However, this new unidimensional raw score may include item scores that either measure multiple dimensions, or are affected by nuisance variance more than other items, or measure nothing at all, thus reducing the psychometric qualities of the sum score, which equally weights all items without consideration. A sufficiently general model, however, could be used to differentially weight item scores when an assessment of the impact of a unidimensional latent variable (as a projection or a geometric average if indeed a multidimensional latent variable underlies the response process) on item-level performance is sought. This can be accomplished, for example, by using item-factor analysis, or by applying an IRT model with an estimated item slope parameter, for example the 2-parameter logistic model (e.g., Lord & Novick, 1968). Obviously, more general models that involve multidimensional generalizations of IRT (von Davier, 2008; Haberman, von Davier & Lee, 2008) could be used in these cases to choose between multidimensional and unidimensional approaches. This is, however, only possible if item-level data are analyzed, and is impossible if item parcels are used. As the present paper shows, it is self-evident that certain undesirable effects of multidimensionality or sources of nuisance variables, while salient in the full multidimensional table, simply cannot be detected in a dataset that is based on item parcels.

3. Item parcels are claimed to reduce the idiosyncratic characteristics of items. Although this might be true, it is our contention that the use of parcels often camouflages potentially serious sources of misfit that can bias interpretation of the results (like cross-loadings, secondary factors, method effects, etc.). Indeed, if a source of variance is truly idiosyncratic to a single item, then it will be appropriately modeled as item uniqueness without resorting to the use of parcels. More generally, the claim that sources of idiosyncratic variance at the item level are unimportant should be made explicit, incorporated into the tested models, and evaluated in terms of the impact that they have on interpretation of the results—as demonstrated in the present investigation.

4. Perhaps the most important advantage of the use of parcels is reducing the number of indicators and parameter estimates, particularly when sample sizes are small and the a priori models are complex (e.g. mixture models, multilevel models, etc.). However, many rules of thumb about the ratios of sample sizes to items or estimated parameters for regular SEMs, CFAs and EFAs are inappropriate. Simulation work by Marsh et al. (1998), Ding, Velicer and Harlow (1995; Velicer & Fava, 1998), De Winter, Dodou, and Wieringa (2009) and others, makes it clear that in terms of finding a well-defined factor structure it is better to have more items per factor—not fewer—especially when sample sizes are low; in some cases, even EFA models may be properly estimated when the sample sizes are less than $n = 50$ (De Winter et al., 2009). Indeed, Marsh et al. found that there was a compensatory relation between sample size and the number of items per factor. Although it is better to have a large number of items per factor and a large sample size, a large number of items compensated for a small sample size. To some extent, well defined solutions are more likely when the number of data points is larger. Marsh et al. specifically evaluated this issue about the supposed advantages of parcels based on factors with 12 items that were used to form 2, 3, 4 or 6 item parcels. Even when the appropriate conditions for the use of item parcels were perfectly satisfied by their simulated data and sample sizes were modest, they concluded that the item level solutions tended to be somewhat better behaved than corresponding parcel solutions.

5. According to even a moderate pragmatist perspective to parceling (e.g., Little, et al., 2002), the focus should be on relations between constructs rather than on misspecification (e.g., method effects or cross-loadings) at the item level. Although consistent with our empirical pragmatist perspective endorsed here, there is an implicit assumption that misspecification at the item level has little or no effect on

relations among constructs and is appropriately controlled through the use of parcels. However, the results of the present investigation show that this implicit assumption is false in a variety of different situations. Misspecification at the item level typically translates into biased estimates of relations among constructs. Hence, the pragmatist needs to worry about misspecification at the item level and should be reluctant to camouflage it with the use of item parcels without empirical tests of underlying assumptions and practical consequences of their violation.

6. Distributive parceling strategies have been claimed to provide better solutions than other parceling strategies (e.g., Little, et al., 2002; Kishton & Widaman, 1994), particularly from a pragmatist perspective. If it is appropriate to use parcels (i.e., there is no misspecification at the item level), then distributive and homogeneous parceling strategies are likely to result in similar conclusions. Indeed, the logic of the homogeneous parceling strategy is predicated on the assumption that there is misspecification at the item level. However, if there is misspecification at the item level, then a distributive parceling strategy that confounds the misspecification with the formation of parcels is likely to more fully camouflage the misspecification, merely making it more difficult to detect whilst still misrepresenting the data in ways that undermines the pragmatist rationale. Nevertheless, if there is misspecification at the item level, then all parceling strategies are likely to bias the substantive interpretation of the results and should be used with caution – if at all. Hence there is a logical inconsistency in the claim that distributive parceling strategies are superior to homogeneous strategies, and the recognition that all parceling strategies are based on the assumption of the unidimensionality inherent in ICM-CFA models. Because the underlying rationale of the distributive strategy only makes sense when the assumption of unidimensionality is violated, it is always a dubious choice.

More generally, research (e.g., Sass & Smith, 2006; Sterba & MacCallum, 2010) demonstrates that parcel allocation variability in the construction of item parcels can result in meaningfully different results, particularly when sample size is small. Although parcel allocation variability not a focus of the present investigation, this further argues against the use of parcels in that there is no particular justification for the use of one versus another (and in practice this situation will reflect in part violations of unidimensionality assumptions). A rather ill-advised corollary of this finding is that a researcher could use the parceling strategy in ways that would not only camouflage misspecifications, but that would also lead to the

"preferred" solution in a more general sense. If different parceling strategies lead to substantively different results, then misuse in the sense of choosing the one parceling strategy that leads to stronger support for a preferred model is a possibility.

Williams and O'Boyle (2008) emphasized that a primary motivation for the use of parcels was in order to meet traditional criteria of acceptable fit when analyses at the item level failed to do so. Clearly this is an inappropriate justification for the use of item parcels that is rarely discussed by proponents of parceling, but that we suspect that it has an impact of the behavior of applied researchers—even if unwittingly. Hence, an overarching dictate of the present investigation is that the use of parcels is almost never justified a priori, is usually wrong when appropriately tested empirically, and typically leads to potentially substantial distortions in relation to goodness of fit and parameter estimates (factor loadings, relations among factors, and prediction of outcome variables). Ignorance of these potential limitations to the use of item parcels (an extreme pragmatist perspective) should no longer be an excuse for the continued, largely unquestioned use of item parcels in applied research.

What is a realistic test for the justification of the use of item parcels? Based on the results of the present investigation, the comparison of ESEM and ICM-CFA solutions provides a practical test for the appropriate use of item parcels—at least when the factor structure is sufficiently well understood that the applied researcher knows the correct number of factors. Under these circumstances, item parcels are justified if ESEM and ICM-CFA models are both able to fit the data, the fit of the ESEM model is not significantly better than the ICM-CFA model, and the relations between constructs are similar in the two models. Unless there is at least moderate support for this test, the use of item parcels in ICM-CFA analyses is not recommended. Furthermore, even when the appropriate factor structure is not known and the applied researcher fits the wrong model (as in Study 4), a careful evaluation of correlated residuals is likely to identify problems in the factor structure so that ESEM and ICM-CFA solutions can be compared appropriately. However, if the applied researcher is not even clear about the appropriate number of factors, then the use of item parcels (and even ICM-CFA models) is probably not appropriate.

Limitations of the present investigation: When 'almost never' is 'maybe'

Another popular homily is 'never say never'. Consistent with this good advice, we qualify our claim, suggesting that it is (almost) never appropriate to use parcels. In this section we discuss situations in

which the analysis of parcels might be justified. Of course, the use of parcels is not always wrong: For data simulated from a pure ICM-CFA population generating model with large samples sizes that minimizes sampling variability, results from all the parcel models were accurate in terms of goodness of fit and in capturing the true population correlation. As demonstrated in Study 3, under these conditions the use of item parcels is justified – perhaps even preferable. However, based on our experience and on the previous research reviewed here we suggest that real data almost never conforms to a pure ICM-CFA model. The requirements for the appropriate use of parcels are well known: there must be good empirical evidence to support the posited ICM-CFA factor structure (i.e., there are no cross-loadings, no correlated uniquenesses, no secondary factors, and no other sources of misspecification). Whilst support for this extremely stringent set of requirements is frequently given token lip-service in justifying the use of parcels or ignored altogether (an extreme pragmatist perspective; also see discussion by Williams & O'Boyle, 2008), results here demonstrate that even minor violations of these conditions can have substantial impact in terms of biasing relations among the latent factors. From an empirical pragmatist perspective, it is critical to demonstrate empirically that the mis-fit associated with the use of an ICM-CFA model and the use of parcels is small and substantively unimportant.

Reise's (2012) recent 'rediscovery' of bifactor measurement models provides an alternative but complimentary perspective on unidimensionality issues. Like us, he argues that many psychological measures are not unidimensional, but argues that bifactor models provide a good basis for partitioning variance into general versus group factors, and determining the degree to which item response data are unidimensional. This is a particularly attractive alternative to the homogenous approach to parceling discussed earlier. Based on IRT research, Reise described a plethora of procedures for evaluating when non-undimensional factors are 'unidimensional enough' but also proposed new indices specific to bifactor models. Implicit in the traditional bifactor model is the assumption that each item loads on the general factor and only one group factor, and that the number and structure of group factors are well-know; an extension of the ICM-CFA logic described here. However, Reise also noted that exploratory bifactor modeling is greatly underused in applied research, and provided preliminary support for an exploratory bifactor model with target rotation that allows items to load on multiple group factors. Although beyond the scope of the present investigation, the extension of models considered here to include bifactor

structures (and bifactor models to more fully incorporate ESEM) are potentially important areas of further research. Importantly, the bifactor approach – as in the present investigation – requires researchers to focus on more realistic factor structures at the level of individual items when unidimensionality is violated.

In some applications, it might be reasonable to argue that unidimensionality is not a relevant concern. Thus, for example, Marsh, Wen and Hau (2004; Marsh, Wen, Nagengast & Hau, in press) proposed a multiple indicator approach to the use of latent interactions in which cross-products of indicators associated with each interacting factor are formed to assess the latent interaction. If the number of indicators for each interacting factor is the same, then each cross-product indicator of the latent interaction is based on a single indicator from each of the interacting factors. However, for example, if one interacting factor is based on 6 items and the other 3, they suggested that it might be appropriate to construct cross-products using 3 item-pairs parcels for the first factor paired with three items for the second factor. Importantly, both the first-order ('main effect') factors are based on items, and it is only the latent interaction factor that is based in part on item parcels. Implicit in this strategy is the assumption that it is the unidimensionality of the first-order factors that is critical rather than the interaction factor. Indeed, alternative distribution-analytic and Bayesian approaches to evaluating latent interactions (Marsh, Wen, Nagengast & Hau, in press; Klein & Muthén, 2007) do not even use multiple indicators of the latent interaction.

A more difficult issue is the use of parcels to obtain the optimal balance between parsimony, accuracy, and bias. Thus, for example, Lüdtke, et al. (2011) as well as many others have shown that for complex models with small Ns, a slightly biased, more parsimonious model might be more accurate than an unbiased, less parsimonious model. That is, the uncertainty region associated estimates based on an unbiased model might be so great that the average deviation between the true and observed model is actually greater than that for the biased model. Thus, for complex models with small Ns, the use of parcels might be justified if the use of parcels substantially reduces the standard errors of the estimates but has no substantively meaningful influence on the size and direction of effects; but it is difficult to know whether the estimates are trustworthy. Indeed, when N is small, tests of statistical significance might not have sufficient power to reject potentially serious violations of unidimensionality (or provide support for associated ESEM models). However, it is precisely under these conditions that lead to high parcel

allocation variability (Sterba & MacCallum, 2010; Sterba, 2011) such that apparently benign differences of the way parcels are constructed can result in apparently important differences in the substantive interpretation of results. More importantly, there are new and evolving Bayesian statistical procedures especially designed for the evaluation of complex factor structures with small Ns where maximum-likelihood might not be appropriate. Thus, for example, the BSEM procedure in Mplus fits a factor model in which cross-loadings and correlated uniquenesses can take on non-zero values with informative priors based on the researcher's judgment. As emphasized by Muthén and Asparouhov (2012), this BSEM rationale is similar in many ways to the target rotation with ESEM demonstrated here, but apparently overcomes potential limitations of ESEM particularly when the model is large relative to the sample size. Hence the primary justification for the use of parcels is likely to be superseded with further development of BSEM. Although BSEM might also supersede ESEM, we view the two approaches as complimentary in which increasing knowledge based on ESEM provides a basis for specifying priors in BSEM, but further research is needed juxtapositions these approaches (for further discussion, see Muthén & Asparouhov, 2012). Nevertheless, particularly when N is small, BSEM estimates are heavily dependent on the analyst's beliefs such that informative priors do not allow the estimates to differ substantially from expected values so that BSEM is not a panacea under these circumstances.

In the present investigation, we have not given much attention to parcel allocation variability (Sterba & MacCallum, 2010; Sterba, 2011), but this is a critical and worrisome issue. However, we have taken a strong stand that distributive parceling strategies that specifically confound sources of misfit and allocation of items to parcels are typically inappropriate, particularly when they provide noticeably better fitting solutions than those based on homogeneous strategies. Although distributive strategies might be explicit, more typically they are implicit through the use of sequential or random assignment of items to parcels. Parcel allocation variability will be most substantial when violations of unidimensionality are substantial. Indeed, it is only when violations of unidimensionality assumptions are substantial (and the use of parcels is most problematic) that distributive and homogeneous strategies are likely to differ systematically. By using only homogeneous parceling strategies in favor of (explicit or implicit) distributive strategies, parcel allocation variability should be reduced substantially. However, even when unidimesionality holds in the population, it will typically not hold in the sample (Sterba, 2011; Sterba &

MacCallum), particularly when sampling variability is substantial (e.g., sample size is small). Hence, more research is needed on how most appropriately to operationalize homogeneous parceling strategies that minimize parcel allocation variability even when the use of parcels might otherwise be justified.

In addition to these issues, there are other limitations in the present investigation that require additional consideration or serve as directions for further research. Analyses presented here are based on ML and MLR estimation, but in some situations (e.g., Likert response scales with less than 4 categories or binary data), alternative estimation procedures might be appropriate. Consistent with results by Bandalos (2008), we contend that issues raised here apply to categorical data as well. Indeed, the use of item parcels might be even more problematic with categorical data (Bandalos, 2008; also see Von Davier, 2010, and earlier discussion).

We recognize that there is necessarily a degree of subjectivity in the interpretation of whether bias is substantively important and that this interpretation might depend on the particular application. Thus, for example, our main focus in Study 1 was the appropriate modeling of method effects which we showed was undermined by the use of item parcels. Conversely, it might be argued that differences in the sizes of latent correlations in Study 1 were not substantively affected, so that the use of item parcels was justified even though parceling confounded method effects and inflated goodness of fit. However, such method effects can bias interpretations in some circumstances. For example, Marsh (1986) showed that for academic self-concept response by young children, negative-item method effects like those considered in Study 1 were substantially related to age and verbal ability within age cohorts; saying false to negatively-worded items to reflect a positive self-concept was more cognitively demanding than responding to positively worded items. This negative-item bias led to inflated estimates of the validity of self-concept responses in relation to academic achievement and complicated interpretations of age-related changes in self-concept. These pragmatically important implications of method effects might have been lost if item parcels were used in a way the camouflaged this source of misfit.

More generally, we do not argue that all sources of misspecification are equally important and that none can be ignored. Rather, our concern is that it is incumbent on the applied researcher to provide substantive, theoretical, and empirical justification for the use of parcels in relation to the substantive aims of their research so that the implications are open to critical scrutiny.

Areas of Agreement/Disagreement in Pro- vs. Con-Parcel Advocates.

In conclusion, it is relevant to contrast our empirical pragmatist (anti-parceling) position with a pragmatist (pro-parceling) position such as that in the influential article by Little et al. (2002). Indeed, there are many areas of agreement as well a few key areas of disagreement. Little et al. (2002), like us, argues from domain-sampling and data-aggregation perspectives that the use of more items is better than the use of fewer items. Although the use of more items that has prompted consideration of item parcels, neither of these perspectives provides justification for the use of parcels rather than items. Both groups acknowledged that the use of parcels is inappropriate when the focus is on scale construction, although we extended this to include related issues of differential item function that are central to studies of latent mean differences (over groups or occasions) and even the justification of scale scores based on manifest means. Like Little et al., we acknowledge that the use of parcels is justified under certain circumstances, but for us these circumstances are limited primarily to models where there is support for unidemensionality at the item level. Although both groups discussed different parceling strategies, Little et al seemed to endorse distributive parceling strategies that confound sources of misfit with the construction of item parcels. In contrast, we argued distributive strategies are typically the least appropriate– apparently for some of the same reasons why they appeal to Little et al. If parcels are to be used at all when unidimensional assumptions are violated, we prefer the homogenous strategies that at least make the sources of misspecification more explicit and open to public scrutiny (but suggest that a bifactor model might be more appropriate). Little et al. posed a distinction between empiricists (who are interested in the psychometric structure of constructs) and pragmatists (who are more interested in relations among constructs). However, to the extent that being a pragmatist relieves applied researchers of responsibility to understand inevitable sources of misspecification at the item level and to provide empirical justification for the use of parcels (i.e., an extreme pragmatist position), we see this as a dangerous distinction. Indeed, whereas we focus on empirical tests of when parceling is or is not appropriate, Little et al. provides little in the way of testable criteria to justify the use of parcels. Nevertheless, from our empirical pragmatist perspective we endorse Little, et al's (2002) conclusion that "the unconsidered use of parcels in never warranted" (p. 151) and their recommendation that "investigators acquire a thorough understanding of the nature and dimensionality of the items to be parceled" (p. 151). Similarly, we argue that the use of parcels is never appropriate a priori

(an extreme pragmatist perspective) and must be justified on the basis of theory, prior research, and particularly preliminary analyses based on the data to be used in the study. However, we probably place more onus on the researcher to make explicit the empirical justification for the use of parcels and implications of misspecification of in the item-level ICM-CFAs that are camouflaged by the use of parcels. Perhaps our most important area of disagreement is the Little et al implicit assumption that sources of misspecification at the item level camouflaged by the use of parcels have no implications for relations among variables and prediction of outcomes. This seems to be the primary basis of their pragmatist perspective and, perhaps, their endorsement of the distributive strategies. In contrast, we have demonstrated that this assumption is frequently false and argue that it should always be made explicit and tested empirically. Pragmatists apparently prefer to ignore problems at the item level, leading them to accept models based on parcel data that would be rejected at the item level without necessarily understanding the consequences. In contrast our empirical pragmatist perspective is that we would rather know about problems at the item level, understand their consequences, and develop more appropriate strategies (or measures) to deal with them in ways that would not be possible at the parcel level. We assume that Little et al would join us in arguing against an extreme pragmatist perspective that is still common place in applied research, so that the main are of disagreement is what empirical support is needed to justify the use of parcels when assumptions underlying their use are violated.

Conclusions and Recommendations

1. Avoid using parcels to camouflage method effects, cross-loadings, and other sources of misspecification at the item level. As demonstrated here, it is better to systematically model the misspecification at the item level. Failure to do so might turn out to be substantively important and will typically bias interpretations of substantively important parameter estimates.

2. The use of item parcels is only justified when there is good support for the unidimensionality of all the constructs at the item level for the particular models and sample being considered. Tests for this requirement should be conducted for the complete model at the item level—not simply the evaluation of each construct separately, which is likely to ignore many forms of misspecification (e.g., cross-loadings, method effects that are common to different constructs). As demonstrated here, a useful test of this requirement is the comparison of a priori ICM-CFA models and corresponding ESEM models. Item

parcels are justified when goodness of fit and parameter estimates based on the more parsimonious ICM-CFA model are good, and similar to those based on the corresponding ESEM. If neither ICM-CFA nor ESEM models fit the data, explore alternative (ex-post facto) solutions at the item level with appropriate caution. Especially in this case, the use of item parcels is likely to be inappropriate. However, we do not argue that the comparison of ESEM and ICM-CFA solutions is the only way to evaluate the appropriateness of item parcels. Alternatives might include application of evolving Bayesian estimation strategies or assessing variation in results associated with different parceling allocations (e.g., Sterba & MacCallum, 2010; Sterba, 2011). For us, the critical issue is that there should appropriate tests of the unidimensionality assumption at the level of the item based on the data actually used in the study and of implications of its violation to interpretations of results based on parcel data.

3. If parcels are used, then applied researchers should use homogeneous rather than distributive parcel strategies that confound sources of misfit with allocation of items to parcels. They should make explicit the parceling strategy used, its justification, and – particularly if sampling variability is large (e.g., sample sizes are small) – whether substantively meaningful differences are associated with different parceling strategies (see Sterba, 2011; Sterba & MacCallum, 2010). Justification for the use of item parcels should be accompanied by at least a summary of the corresponding results based on item analyses and any substantively important differences between the two. However, if high parcel allocation variability results in ambiguous interpretations based on parcels or there are substantively important differences between item and parcel results, then applied researchers are advised to use item level analyses.

4. Applied researchers will continue to argue, sometimes with good justification, that there are situations (e.g., very large, complex SEMs with many indicators and small sample sizes) in which it is not reasonable to replicate fully the analyses based on parcels with models based on items. However, even here we believe that, especially in light of the studies cited earlier on sample size requirements, it is reasonable to conduct preliminary tests of unidimensionality at the item level. Thus, for example, researchers might adapt McDonald's (2010) recommendation to use a two-step approach in which tests of the measurement model are conducted in the first step; ICM-CFA and ESEM comparisons could be based on the measurement model. Then, if there is good support for unidimensionality, researchers could proceed to the use of item parcels in the more complex SEMs in the second step. If even the full measurement model is

too complex to fit at the item level, researchers might evaluate the factor structure of logically defined subsets of factors in relation to different subsets of factors (e.g., the multiple factors based on with instrument in relation to multiple factors based on each of the other instruments in a pairwise strategy). Also, evolving Bayesian estimation procedures might make large models more tractable. Nevertheless, the onus is still on the researcher to justify the assumption of unidimensionality of their measures as a basis for using item parcels.

In conclusion, it is almost always preferable to evaluate latent variable models based on item-level data, particularly when the sample size is appropriate. A priori use of item parcels (an extreme pragmatist perspective) is never justified without clear support for the a priori model at the item level and unidimensionality in relation to the models and data under consideration. Expedient compromises between parsimony and accuracy in applied research when sample sizes are modest (e.g., the use of parcel scores) are likely to be biased under typical conditions. They should be avoided unless there is clear evidence that the very restrictive unidimensionality assumptions upon which they are based are met, or that the sizes of biases are trivially small and substantively unimportant. Parcels are widely over-used for the wrong reasons and are only defensible under an extremely narrow set of circumstances that are rarely considered (and unlikely to be met) in most studies that use them. Although we do not argue that it is absolutely never appropriate to use item parcels, we argue that their use is typically inappropriate and often based on the wrong reasons. In this sense the overarching purpose of this article is to raise the consciousness of journal editors, reviewers, and consumers of research – as well as applied researchers themselves – about the widespread misuse of item parcels, criteria needed to justify their use, and potential implications of the failure to heed these warnings.

Footnotes

1. Mplus currently does not have the option of using a covariance matrix as input for ESEM. Instead, we used a very large sample (N = 100,000) to approximate a population. We used this strategy in that our focus was on misspecification at the population level rather than on sampling per se.

References

Alhija, F. N. A., & Wisenbaker, J. (2009). A Monte Carlo study investigating the impact of item parceling strategies on parameter estimates and their standard errors in CFA. *Structural Equation Modeling, 13,* 204-228.

Anderson, S. E., Coffey, B. S., & Byerly, R. T. (2002). Formal Organizational Initiatives and Informal Workplace Practices: Links to Work-Family Conflict and Job-Related Outcomes. *Journal of Management 28,* 787-810.

Asparouhov, T. & Muthén, B. (2009). Exploratory structural equation modeling. *Structural Equation Modeling, 16,* 397-438.

Bachman, J. G. (2002). Volume I of the Documentation Manual. Ann Arbor, MI: Inter-university Consortium for Political and Social Research.

Bagozzi, R. P., & Edwards, J. R. (1998). A general approach for representing constructs in organizational research. *Organizational Research Methods*, *1*, 45–87.

Bandalos, D. L. (2002). The effects of item parceling on goodness-of-fit and parameter estimate bias in structural equation modeling. *Structural Equation Modeling*, *9*, 78–102.

Bandalos, D. L. (2008). Is parceling really necessary? A comparison of results from item parceling and categorical variable methodology. *Structural Equation Modeling*, *15*, 211–240.

Bandalos, D. L., & Finney, S. J. (2001). Item parceling issues in structural equation modeling. In G. A. Marcoulides & R. E. Schumacker (Eds.), *New developments and techniques in structural equation modeling* (pp. 269–296). Mahwah, NJ: Lawrence Erlbaum.

Borkenau, P., & Ostendorf, F. (1990). Comparing exploratory and confirmatory factor analysis: A study on the 5-factor model of personality. *Personality and Individual Differences, 11,* 515–524.

Box, G. E. P. (1979). Some problems of statistics and everyday life. *Journal of the American Statistical Association*, *74*, 1–4.

Browne, M. W. (2001). An overview of analytic rotation in exploratory factor analysis. *Multivariate Behavioral Research*, *36*, 111–150.

Cattell, R. B. (1949). A note on factor invariance and the identification of factors. *British Journal of Psychology*, *2*, 134–139.

Cattell, R. B. (1956). Validation and intensification of the sixteen personality factor questionnaire. *Journal of*

Clinical Psychology, *12*, 205–214.

Cattell, R. B. (1974). Radial parcel factoring vs. item factoring in defining personality structure in questionnaires: Theory and experimental checks. *Australian Journal of Psychology*, *26*, 103–119.

Comrey, A. L. (1984). Comparison of two methods to identify major personality factors. *Applied Psychological Measurement*, *8*, 397–408.

Coffman, D. L. & MacCallum, R. C. (2005). Using parcels to convert path analysis models into latent variable models. Multivariate Behavioral Research, 40, 235-259.

Cook, K. F., Kallen, M. A., & Amtmann, D. (2009). Having a fit: Impact of number of items and distribution of data on traditional criteria for assessing IRT's unidimensionality assumption. *Quality of Life Research, 4*, 447–460.

Costa, P. T., & McCrae, R. R. (1992). *Revised NEO Personality Inventory (NEO PI-R) and NEO Five Factor Inventory (NEO-FFI): Professional manual*. Odessa, FL: Psychological Assessment Resources.

Croon, M. (2002). Using predicted latent scores in general latent structure models. In G. A. Marcoulides & I. Moustaki (Eds.), Latent variable and latent structure models (pp. 195–223). Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

Cudeck, R. & Henly, S. J. (1991). Model selection in covariance structures analysis and the "problem" of sample size: A clarification. *Psychological Bulletin*, *109*, 512–519.

De Winter, J. C. F., Dodou, D., & Wieringa, P. A. (2009). Exploratory factor analysis with small sample sizes. *Multivariate Behavioral Research*, *44*, 147–181.

Ding, L., Velicer, W. F., & Harlow, L. L. (1995). The effects of estimation methods, number of indicators per factor and improper solutions on structural equation modeling fit indices. *Structural Equation Modeling*, *2*, 119–144.

Dolan, C. V., Oort, F. J., Stoel, R. D., & Wicherts, J. M. (2009). Testing measurement invariance in the target rotated multigroup exploratory factor model. *Structural Equation Modeling*, *16*, 295–314.

Enders, C.K., & Bandalos, D. L. (2001), The relative performance of full information maximum likelihood estimation for missing data in structural equation models. Structural Equation Modeling, 8, 430-457.

Gallucci, M., & Perugini, M. (2007). The Marker Index: A new method of selection of marker variables in factor analysis. *TPM-Testing, Psychometrics, Methodology in Applied Psychology*, *14*, 3–25.

Haberman, S. J., von Davier, M., & Lee, Y. (2008). Comparison of Multidimensional Item Response

Models: Multivariate Normal Ability Distributions Versus Multivariate Polytomous Ability Distributions. *RR-08-45. ETS Research Report.*

Hall, R. J., Snell, A. F., & Foust, M. S. (1999). Item parceling strategies in SEM: Investigating the subtle effects of unmodeled secondary constructs. *Organizational Research Methods, 2*, 233-256.

Hau, K.-T., & Marsh, H. W. (2004). The use of item parcels in structural equation modeling: Nonnormal data and small sample sizes. *British Journal of Mathematical and Statistical Psychology, 57*, 327–351.

Holden, R. R., & Fekken, G. C. (1994). The NEO Five-Factor Inventory in a Canadian context: psychometric properties for a sample of university women. *Personality and Individual Differences, 17*, 441–444.

Hopwood, C. J., & Donnellan, M. B. (2010) How should the internal structure of personality inventories be evaluated? *Personality and Social Psychology Review, 14*, 3, 332–346.

Howarth, E. (1972). A factor analysis of selected markers for objective personality factors. *Multivariate Behavioral Research, 7(4)*, 451–476.

Hu, L., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling, 6*, 1–55.

Jöreskog, K. G. (1979). Statistical estimation of structural models in longitudinal investigations. In J. R. Nesselroade & B. Baltes (Eds.), *Longitudinal research in the study of behavior and development* (pp. 303–351). New York: Academic Press.

Kishton, J. M., & Widaman, K. F. (1994). Unidimensional versus domain representative parceling of questionnaire items: An empirical example. *Educational and Psychological Measurement, 54*, 757-765.

Klein, A. G., & Muthén, B. O. (2007). Quasi-maximum likelihood estimation of strucural equation models with multiple interaction and quadratic effects. *Multivariate Behavioral Research, 42*, 647-663.

Little, T. D., Cunningham, W. A., Shahar, G., & Widaman, K. F. (2002). To parcel or not to parcel: Exploring the question, weighing the merits. *Structural Equation Modeling, 9*, 151–173.

Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores*. Reading, MA: Addison-Wesley.

MacCallum, R. C. (2003). 2001 Presidential Address: Working with Imperfect Models', Multivariate Behavioral Research, 38: 1, 113–139.

MacCallum, R.C. & Tucker, L.R (1991). Representing sources of error in the common-factor model: Implications for theory and practice. *Psychological Bulletin, 109*, 502-511.

MacCallum, R. C.,Widaman, K. F., Zhang, S., & Hong, S. (1999). Sample size in factor analysis. *Psychological Methods, 4,* 84–99.

Marsh, H. W. (1986). The bias of negatively worded items in rating scales for young children: A cognitive-developmental phenomena. Developmental Psychology, 22, 37-49.

Marsh, H. W. (2007). Application of confirmatory factor analysis and structural equation modeling in sport/exercise psychology. In G. Tenenbaum & R. C. Eklund (Eds.), Handbook of sport psychology (3rd ed., pp. 774–798). New York: Wiley.

Marsh, H. W., Balla, J. R., & Hau, K. T. (1996). An evaluation of incremental fit indexes: A clarification of mathematical and empirical processes. In G. A. Marcoulides & R. E. Schumacker (Eds.), *Advanced structural equation modeling techniques* (pp. 315–353). Hillsdale, NJ: Erlbaum.

Marsh, H. W., Balla, J. R., & McDonald, R. P. (1988). Goodness-of-fit indexes in confirmatory factor analysis: The effect of sample size. *Psychological Bulletin, 103,* 391–410.

Marsh, H. W., & Hau, K-T. (1996). Assessing goodness of fit: Is parsimony always desirable? *Journal of Experimental Education, 64*, 364–390.

Marsh, H. W., Hau, K-T., Balla, J. R., & Grayson, D. (1998). Is more ever too much? The number of indicators per factor in confirmatory factor analysis. *Multivariate Behavioral Research, 33,* 181–220.

Marsh, H. W., Hau, K-T., & Grayson, D. (2005). Goodness of Fit Evaluation in Structural Equation Modeling. In A. Maydeu-Olivares & J. McCardle (Eds.). *Psychometrics. A Festschrift to Roderick P. McDonald*. Hillsdale, NJ: Erlbaum.

Marsh, H. W., Hau, K.T., & Wen, Z., (2004). In search of golden rules: Comment on hypothesis testing approaches to setting cutoff values for fit indexes and dangers in overgeneralising Hu & Bentler's (1999) findings. *Structural Equation Modeling, 11,* 320–341.

Marsh, H. W., Liem, G. A. D., Martin, A. J., Morin, A. J. S., & Nagengast, B. (2011a). Methodological measurement fruitfulness of exploratory structural equation modeling (ESEM): New approaches to key substantive issues in motivation and engagement. *Journal of Psychoeducational Assessment, 29,* 322-346.

Marsh, H. W., Lüdtke, O., Muthén, B., Asparouhov, T., Morin, A. J. S., Trautwein, U., & Nagengast, B. (2010). A new look at the big-five factor structure through exploratory structural equation modeling. *Psychological Assessment*, 22, 471-491.

Marsh, H. W., Muthén, B., Asparouhov. T., Lüdtke, O., Robitzsch, A., Morin, A.J.S., & Trautwein, U. (2009). Exploratory Structural Equation Modeling, Integrating CFA and EFA: Application to Students' Evaluations of University Teaching. *Structural Equation Modeling, 16,* 439–476.

Marsh, H.W., Nagengast, B., Morin, A.J.S., Parada, R.H., Craven, R.G., & Hamilton, L.R. (2011b). Construct Validity of the Multidimensional Structure of Bullying and Victimization: An Application of Exploratory Structural Equation Modeling. *Journal of Educational Psychology, 103,* 701-732.

Marsh, H. W. & O'Neill, R. (1984). Self Description Questionnaire III (SDQ III): The construct validity of multidimensional self-concept ratings by late-adolescents. *Journal of Educational Measurement, 21,* 153–174.

Marsh, H. W., Scalas, L. F., & Nagengast, B. (2010). Longitudinal tests of competing factor structures for the Rosenberg Self-Esteem Scale: Traits, ephemeral artifacts, and stable response styles. *Psychological Assessment, 22,* 366-381.

Marsh, H. W.; Wen, Z.; Hau, K. (2004). Structural equation models of latent interactions: Evaluation of alternative estimation strategies and indicator construction. *Psych Methods, 9, 275-300.*

Marsh, H. W., Wen, Z., Nagengast, B. & Hau, K-T. (in press). Structural equation models of latent interaction. In Hoyle, R. & West, S. (Eds). Handbook of Structural Equation Modeling. Guilford Press, NY.

McCrae, R. R., & Costa, P. T. Jr. (2004). A contemplated revision of the NEO Five-Factor Inventory. *Personality and Individual Differences, 36,* 587–596.

McCrae, R. R., Zonderman, A. B., Costa, P. T., Jr., Bond, M. H., & Paunonen, S. (1996). Evaluating the replicability of factors in the revised NEO Personality Inventory: Confirmatory factor analysis versus procrustes rotation. *Journal of Personality and Social Psychology, 70,* 552–566.

McDonald, R. P. (2010). Structural models and the art of approximation. *Perspectives on Psychological Science, 5,* 675-686.

Meade, A. W., & Kroustalis, C. M. (2006). Problems with item parceling for confirmatory factor analytic tests of measurement invariance. *Organizational Research Methods, 9,* 369–403.

Meredith, W. (1993). Measurement invariance, factor analysis and factorial invariance, *Psychometrika, 58,* 525–543.

Meredith, W., & Teresi, J. (2006). An essay on measurement and factorial invariance. *Medical Care, 44*

(Suppl 3), S69–S77.

Morin, A. J. S., Marsh, H. W. & Nagengast, B. (in press). Exploratory structural equation models. In Hoyle, R. & West, S. (Eds). Handbook of Structural Equation Modeling. Guilford Press, NY.

Mulaik, S. A. (1972). *The foundations of factor analysis*. New York: McGraw-Hill.

Muthén, B. & Asparouhov, T. (2012). Bayesian structural equation modeling: A more flexible representation of substantive theory. Psychological Methods. 17, 313-335. DOI: 10.1037/a0026802

Muthén, L. K., & Muthén, B. (2010). *Mplus user's guide.* Los Angeles CA: Muthén & Muthén.

Nunnally, J. C. (1978). Psychometric theory (2nd ed.). New York: McGraw-Hill.

Overall, J. E. (1974). Marker variable factor analysis: A regional principal axes solution. *Multivariate Behavioral Research, 9*(2), 149–164.

Reise, S. P. (2012). The rediscovery of bifactor measurement models, multivariate behavioral research, 47, 667-696

Rogers, W. M., & Schmitt, N. (2004). Parameter Recovery and Model Fit Using Multidimensional Composites: A Comparison of Four Empirical Parceling Algorithms. *Multivariate Behavioral Research, 39*, 379-412.

Sass, D. A., & Smith, P. L. (2006). The effects of parceling unidimensional scales on structural parameter estimates in structural equation modeling. *Structural Equation Modeling, 13*, 566–586.

Sass, D. A. and Schmitt, T. A. (2010). A comparative investigation of rotation criteria within exploratory factor analysis. *Multivariate Behavioral Research, 45*, 73–103.

Schmitt, T. A., & Sass, D. A. (2011). Rotation criteria and hypothesis testing for exploratory factor analysis: Implications for factor pattern loadings and interfactor correlations. *Educational and Psychological Measurement, 71*, 95-113.

Sterba, S. K., & Sterba, S.K. (2011). Implications of parcel-allocation variability for comparing fit of item-solutions and parcel-solutions. *Structural Equation Modeling, 18,* 554-577.

MacCallum, R. C. (2010). Variability in parameter estimates and model fit across repeated allocations of items to parcels. *Multivariate Behavioral Research, 45*(2), 322–358.

Thissen, D. (2001). Psychometric engineering as art. *Psychometrika, 66*, 473–486.

Thurstone, L. L. (1930). The learning function. Journal of General Psychology, *3*, 469–478.

Tukey, J. W. (1961). Discussion, emphasizing the connection between analysis of variance and spectrum

analysis. *Technometrics*, *3*, 191–219.

Velicer, W. F., & Fava, J. L. (1998). Effects of variable and subject sampling on factor pattern recovery. *Psychological Methods*, *3*, 231–251.

von Davier, M. (2008). A general diagnostic model applied to language testing data. *British Journal of Mathematical and Statistical Psychology*, *61*, 287–307.

von Davier, M. (2010). Why sum scores may not tell us all about test takers. In Wang, L. (Ed.): *Special issue on Quantitative Research Methodology. Newborn and Infant Nursing Reviews*. *10*, 27–36.

Williams, L. J., & O'Boyle, E. H., Jr. (2008). Measurement models for linking latent variables and indicators: A review of human resource management research using parcels. *Human Resource Management Review, 18,* 233-242.

Yang, C., Nay, S., & Hoyle, R.H. (2010). Three Approaches to Using Lengthy Ordinal Scales in Structural Equation Models: Parceling, Latent Scoring, and Shortening Scales. *Applied Psychological Measurement*, *34*(2) 122–142.

Table 1

Study 1: Confirmatory Factor Analysis and Tests of Invariance For Measures of Self-Esteem on Four Occasions

Model	χ^2	*df*	*TLI*	*CFI*	***RMSEA***
One Self-Esteem Factor, Longitudinal tests of invariance —no method effects					
1a Unconstrained model					
10 Items	2930.10	674	.838	.860	.039
5 Parcels	230.79	134	.987	.991	.018
3 Parcels	17.27	30	1.004	1.000	.000
3 Parcels (homo)	104.29	30	.977	.989	.033
1b Factor Loadings (FL)					
10 Items	2971.55	701	.843	.859	.038
5 Parcels	241.64	146	.988	.991	.017
3 Parcels	21.73	36	1.003	1.000	.000
3 Parcels (homo)	144.95	36	.971	.984	.037
1c FL & Variances (Var)					
10 Items	2975.17	704	.844	.859	.038
5 Parcels	243.87	149	.988	.991	.017
3 Parcels	24.68	39	1.003	1.000	.000
3 Parcels (homo)	147.49	39	.974	.985	.035
1d FL-Var-Uniqueness (UNQ)					
10 Items	3134.95	724	.839	.850	.039
5 Parcels	519.05	190	.961	.966	.031
3 Parcels	204.56	48	.973	.980	.038
3 Parcels (homo)	367.03	48	.937	.954	.055
One Self-Esteem Factor, Positive & Negative Item Method Factors, Longitudinal Tests of Invariance					
2-0 Method Factors uncorrelated over time [a]	1483.37	634	.935	.947	.025
2a Unconstrained	916.52	622	.977	.982	.015
2b FFL	962.06	673	.979	.982	.014
2c FL & Var	1000.90	682	.977	.980	.015
2d FL-Var-Uniq	1161.90	702	.968	.971	.017

Note. See Figure 1 for a description of the models. Separate analyses were done for each wave separately (see results in Appendix 2, supplemental materials) and then longitudinal analyses across all four waves. For the longitudinal analyses presented here, separate tests were done to evaluate the invariance of factor loadings (FL), factor variances (VAR), and uniquenesses (UNQ). The 5 Parcel and 3 Parcel models are based on parcels with a mix of positive and negatively worded items. The 3 Parcels (homo) model is based on one parcel with four negatively worded items and two parcels each containing three positively worded items. χ^2 = chi-square test statistic; df = degrees of freedom; TLI= Tucker–Lewis index; CFI = comparative fit index; RMSEA = root-mean-square error of approximation.

[a]In Model 2-0, all positive and negative item method factors are constrained to be uncorrelated. In subsequent models (2a-2d) positive time method factors are allowed to be correlated over time, as are the negative item method factors (see Table 2). However, positive and negative item method factors are constrained to be uncorrelated with each other in all models.

Table 2

Study 1: Latent Correlations Among Self-Esteem Factors on Four Occasions (T1–T4)

	Self-Esteem				Positive Item Method Factor				Negative Item Method Factor			
	T1	**T2**	**T3**	**T4**	**T1**	**T2**	**T3**	**T4**	**T1**	**T2**	**T3**	**T4**
Model 2: 10 Items: Self-esteem & Method Factors												
T1	**1.0**				**1.0**				**1.0**			
T2	**.71**	**1.0**			**.52**	**1.0**			**.49**	**1.0**		
T3	**.64**	**.82**	**1.0**		**.48**	**.60**	**1.0**		**.39**	**.49**	**1.0**	
T4	**.55**	**.68**	**.80**	**1.0**	**.43**	**.62**	**.60**	**1.0**	**.39**	**.65**	**.60**	**1.0**
Model 1: 10 Items: Self-esteem & No Method Factors												
T1	**1.0**											
T2	**.65**	**1.0**										
T3	**.59**	**.76**	**1.0**									
T4	**.51**	**.66**	**.76**	**1.0**								
Model 1: Five Parcels												
T1	**1.0**											
T2	**.63**	**1.0**										
T3	**.59**	**.77**	**1.0**									
T4	**.52**	**.67**	**.77**	**1.0**								
Model 1: Three Parcels												
T1	**1.0**											
T2	**.65**	**1.0**										
T3	**.60**	**.76**	**1.0**									
T4	**.52**	**.68**	**.78**	**1.0**								
Model 1: Three Parcels (Positive &Negative Parcels)												
T1	**1.0**											
T2	**.63**	**1.0**										
T3	**.59**	**.77**	**1.0**									
T4	**.51**	**.67**	**.77**	**1.0**								

Note. See Figure 1 for a description of the models. T1–T4 = Times 1 to 4.

Table 3

Study 2: Summary of Goodness of Fit Statistics for Models of Neuroticism & Extraversion

Model	CHI	df	CFI	TLI	RMSEA	Md Parm Est		FCorr	Data
						Fac1	Fac2	Unq	
ICM-CFA for Items-two factors									
RDCFA1	2349	231	.866	.840	.052	.54	.43	.75	-.49	24 items
ICM-CFA for Parcels-two factors									
RDCFA2	733	53	.930	.912	.062	.64	.61	.61	-.52	12 parc
RDCFA3	412	19	.949	.925	.078	.71	.71	.50	-.49	8 parc
RDCFA4	44	8	.994	.989	.036	.76	.76	.42	-.47	6 parc
ESEM for Items-Target rotation-two factors									
RDtrg1	1457	209	.921	.896	.042	.51	.39	.75	-.15	24 items
ESEM/CFA for Items—one factor (homogeneous strategy)									
RD1CFA1a	6026	252	.636	.601	.082	.42	--	.70	--	24 items
ESEM/CFA for Parcels— one factor (homogeneous strategy)									
RD1CFA2a	3313	54	.654	.577	.137	.51	---	.66	---	12 parc
RD1CFA3a	2660	20	.657	.520	.197	.67	---	.66	---	8 parc
RD1CFA4a	2098	9	.617	.361	.262	.55	---	.70	---	6 parc
ESEM/CFA for Parcels— one factor (distributive strategy)									
RD1CFA2b	602	54	.937	.923	.055	.75	---	.57	---	12 parc
RD1CFA3b	175	20	.979	.970	.048	.66	---	.43	---	8 parc
RD1CFA4b	137	9	.978	.963	.065	.67	---	.55	---	6 parc

Note. CHI = chi-square; df = degrees of freedom ratio; CFI = comparative fit index; TLI = Tucker-Lewis Index; RMSEA = Root Mean Square Error of Approximation. CFA = confirmatory factor analysis. ESEM = exploratory structural equation modelling. Scale score = unweighted average of items designed to measure each scale; Scale Score corrected = Scale scores corrected for coefficient alpha estimates of reliability; Factor score = factor scores generated with each of the factor analysis models.

[a] . For purposes of convenience of reporting, we reverse-scored all the items so that all items loaded positively on the single latent factor.

Table 4

Study 3: Summary of Goodness of Fit Statistics for Simulated Data

		Population ρ = .25							Population ρ = .60						
Model	df	CFI	TLI	RMSEA	Md Parm Est			Fac	CFI	TLI	RMSEA	Md Parm Est			Fac
					Fac1	Fac2	Unq	Corr				Fac1	Fac2	Unq	Corr
Pure Simple-Structure															
CFA for Items-two factors															
24item	251	1.000	1.000	.000	.60	.60	.64	.25	1.000	1.000	.000	.60	.60	.64	.60
12parc	53	1.000	1.000	.001	.68	.68	.53	.25	1.000	1.000	.000	.68	.68	.53	.60
8parc	19	1.000	1.000	.000	.78	.78	.37	.25	1.000	1.000	.002	.78	.78	.39	.60
6parc	8	1.000	1.000	.001	.82	.82	.34	.25	1.000	1.000	.000	.82	.82	.33	.60
ESEM for Items-Target rotation—two factors															
24items	229	1.000	1.000	.001	.60	.60	.64	.24	1.000	1.000	.001	.60	.60	.64	.60
ESEM/CFA for Items—one factor (unconfounded)															
24item	252	.581	.541	.100	.43	--	.81	(1)	.819	.802	.068	.55	--	.53	(1)
ESEM/CFA for Parcels—one factor (confounded)															
12parc	54	1.000	1.000	.000	.64	---	.59	(1)	1.000	1.000	.001	.69	---	.53	(1)
8parc	20	.974	.964	.062	.69	---	.51	(1)	.994	.991	.035	.74	---	.45	(1)
6parc	9	1.000	1.000	.003	.72	---	.48	(1)	1.000	1.000	.000	.76	---	.42	(1)
Almost Perfect Simple-Structure															
CFA for Items-two factors															
24item	251	.990	.989	.016	.59	.60	.65	.41	.995	.994	.012	.59	.59	.65	.71
12parc	53	.997	.986	.017	.71	.71	.49	.43	.999	.996	.024	.74	.74	.46	.72
8parc	19	1.000	1.000	.003	.79	.79	.37	.43	1.000	1.000	.003	.81	.81	.35	.72
6parc	8	.998	.997	.024	.83	.83	.31	.43	.999	.999	.018	.84	.85	.29	.72
ESEM for Items-Target rotation-two factors															
24items	229	1.000	1.000	.001	.60	.60	.64	.25	.000	1.000	.000	.60	.60	.64	.60
ESEM/CFA for Items—one factor (unconfounded)															
24item	252	.691	.661	.091	.48	--	.77	(1)	.879	.867	.061	.54	--	.71	(1)
12parc	54	.663	.588	.177	.60	---	.60	(1)	.866	.836	.120	.68	---	.53	(1)
8parc	20	.635	.489	.266	.65	---	.61	(1)	.850	.790	.173	.74	---	.44	(1)
6parc	9	.622	.370	.434	.63	---	.55	(1)	.835	.725	.259	.77	---	.41	(1)
ESEM/CFA For Parcels— one factor (confounded)															
12parc	54	1.000	1.000	.001	.64	---	.58	(1)	1.000	1.000	.000	.68	---	.53	(1)
8parc	20	.981	.973	.059	.73	---	.47	(1)	.995	.993	.035	.78	---	.39	(1)
6parc	9	1.000	1.000	.002	.78	---	.39	(1)	1.000	1.000	.000	.82	---	.38	(1)

Table 4 (continued)

		Population ρ = .25							Population ρ = .60						
Model	df	CFI	TLI	RMSEA	Md Parm Est			Fac	CFI	TLI	RMSEA	Md Parm Est			Fac
					Fac1	Fac2	Unq	Corr				Fac1	Fac2	Unq	Corr
Good Simple-Structure															
CFA for Items-two factors															
24item	251	.969	.966	.030	.59	.58	.66	.52	.987	.985	.022	.65	.64	.84	.78
12parc	53	.990	.987	.033	.73	.73	.46	.54	.996	.994	.024	.76	.84	.42	.78
8parc	19	1.000	1.000	.005	.81	.81	.35	.54	1.000	1.000	.004	.82	.82	.32	.78
6parc	8	.996	.992	.041	.84	.84	.30	.54	.998	.997	.030	.86	.86	.26	.78
ESEM for Items-Target rotation-two factors															
24items	229	1.000	1.000	.000	.60	.60	.64	.24	1.000	1.000	.001	.60	.60	.95	.60
ESEM/CFA for Items—one factor (unconfounded)															
24item	252	.754	.730	.085	.55	--	.70	(1)	.908	.899	.058	.62	--	.61	(1)
12parc	54	.738	.680	.163	.60	---	.54	(1)	.899	.876	.112	.71	---	.50	(1)
8parc	20	.710	.594	.248	.68	---	.54	(1)	.886	.840	.173	.77	---	.40	(1)
6parc	9	.691	.486	.434	.69	---	.50	(1)	.875	.792	.241	.80	---	.36	(1)
ESEM/CFA For Parcels— one factor (confounded)															
12parc	54	1.000	1.000	.000	.70	---	.51	(1)	1.000	1.000	.027	.76	---	.43	(1)
8parc	20	.990	.986	.047	.76	---	.42	(1)	.997	.996	.137	.81	---	.34	(1)
6parc	9	1.000	1.000	.002	.80	---	.36	(1)	1.000	1.000	.003	.84	---	.28	(1)
Moderate Simple-Structure															
CFA for Items-two factors															
24item	251	.950	.945	.047	.70	.70	.55	.84	.981	.979	.034	.73	.73	.47	.94
12parc	53	.984	.979	.050	.78	.78	.36	.82	.991	.989	.034	.84	.84	.27	.92
8parc	19	.988	.982	.051	.85	.85	.29	.81	.995	.992	.048	.89	.89	.21	.92
6parc	8	.999	.997	.030	.88	.88	.23	.81	.999	.999	.021	.91	.91	.17	.91
ESEM for Items-Target rotation-two factors															
24items	229	1.000	1.000	.001	.60	.60	.53	.25	1.000	1.000	.001	.60	.60	.46	.60
ESEM/CFA for Items—one factor (homogeneous-unconfounded parcels)															
24item	252	.899	.889	.067	.63	--	.60	(1)	.961	.953	.089	.83	---	.30	(1)
12parc	54	.902	.880	.120	.75	---	.54	(1)	.955	.937	.137	.87	---	.24	(1)
8parc	20	.890	.846	.183	.80	---	.36	(1)	.949	.915	.193	.89	---	.21	(1)
6parc	9	.880	.801	.253	.83	---	.32	(1)							
ESEM/CFA For Parcels— one factor (distributive-confounded parcels)															
12parc	54	1.000	1.000	.000	.77	---	.40	(1)	1.000	1.000	.001	.83	---	.31	(1)
8parc	20	.996	.994	.037	.83	---	.30	(1)	.999	.998	.023	.89	---	.19	(1)
6parc	9	1.000	1.000	.000	.87	---	.23	(1)	1.000	1.000	.002	.92	---	.16	(1)

Note. CHI = chi-square; df = degrees of freedom ratio; CFI = comparative fit index; TLI = Tucker-Lewis Index; RMSEA = Root Mean Square Error of Approximation. CFA = confirmatory factor analysis. ESEM = exploratory structural equation modelling.

Table 5

Study 4: Item and Parcel Solutions Based on Complex Factor Structures

	CFI	TLI	RMSEA	b_1	b_2	b_3	r_{12}	r_{13}	r_{23}	R^2	
Population 1				**.306**	**.306**	**.300**	**.600**	**.300**	**.300**		
Two factor ESEM											
Item Mth	1.000	1.000	.000	.307	.309	.294	.602	.304	.304	.501	
Item No Mth	.995	.994	.015	.304	.307	.294	.598	.309	.306	.501k	
One factor ESEM											
Item Mth	.961	.956	.041	.495		.379	(1)	.296		.500	
Item No Mth	.957	.951	.043	.491		.377	(1)	.304		.496	
Two factor CFA											
Item Mth	.983	.982	.027	.269	.269	.271	.897	.519	.524	.501	
Item No Mth	.977	.975	.031	.253	.281	.269	.903	.533	.529	.496	
Parcel Dist	.986	.982	.051	.274	.282	.254	.876	.488	.487	.492	
Parcel Homo	.971	.962	.070	.271	.264	.269	.902	.537	.532	.497	
One factor CFA											
Item Mth	.954	.950	.044	.525		.268	(1)	.542		.500	P1F1B
Item No Mth	.950	.946	.045	.522		.268	(1)	.545		.497	P1IF1A
Parcel Dist	.986	.982	.054	.538		.254	(1)	.503		.492	
Parcel Homo	.928	.911	.111	.412		.263		.506		.348	
Population 2				**.555**	**.000**	**.309**	**.600**	**.300**	**.300**		
Two factor ESEM											
Item Mth	1.000	1.000	.000	.556	.000	.302	.603	.298	.295	.500	
Item No Mth	.995	.994	.015	.551	-.001	.302	.598	.300	.300	.493	
One factor ESEM											
Item Mth	.958	.953	.043	.440		.379	(1)	.288		.434	
Item No Mth	.955	.949	.044	.448		.373	(1)	.295		.438	
Two factor CFA											
Item Mth	.983	.981	.027	.787	-.307	.282	.898	.519	.520	.500	
Item No Mth	.977	.975	.031	.799	-.323	.280	.904	.529	.529	.496	
Parcel Dist	986	.982	.051	.744	-.246	.266	.877	.483	.487	.492	
Parcel Homo	.970	.962	.071	.832	-.358	.280	.903	.534	.531	.503	
One factor CFA											
Item Mth	.951	.947	.045	.460		.283	(1)	.540		.433	
Item No Mth	.947	.944	.047	.469		.277	(1)	.543		.438	
Parcel Dist	.985	.982	.054	.482		.265	(1)	.500		.430	
Parcel Homo	.938	.923	.101	.470		.276	(1)	.549		.440	

Table 5 continued on next page

Table 5 continued

	CFI	TLI	RMSEA	b_1	b_2	b_3	r_{12}	r_{13}	r_{23}	R^2
Population 3				**.670**	**-.200**	**.300**	**.600**	**.300**	**.300**	
Two factor ESEM										
Item Mth	1.000	1.000	.000	.671	-.197	.298	.599	.299	.299	.504
Item No Mth	.995	.994	.015	.664	-.198	.297	.594	.304	.301	.498
One factor ESEM										
Item Mth	954	.948	.045	.387		.370	(1)	.286		.349
Item No Mth	950	.945	.046	.381		.361	(1)	.294		.356
Two factor CFA										
Item Mth	.982	.981	.027	1.049	-.647	.284	.896	.519	.522	.503
Item No Mth	.977	.975	.031	1.078	-.678	.282	.903	.529	.531	.502
Parcel Dist	.986	.982	.050	.980	-.558	.289	.875	.487	.485	.495
Parcel Homo	.969	.961	.072	1.111	-.714	.283	.900	.535	.532	.517
One factor CFA										
Item Mth	.947	.943	.046	.381		.290	(1)	.542		.348
Item No Mth	.943	.940	.048	.396		.280	(1)	.545		.356
Parcel Dist	.985	.982	.953	.406		.268	(1)	.501		.345
Parcel Homo	.930	.914	.106	.396		.280	(1)	.550		.357

Note. Three population generating models (see Figure 1; also see supplemental materials) differed only in terms of the path coefficients relating the three predictor factors to the outcome variable. Each population generating model had three predictor factors that had a complex structure including cross-loadings and method (MTH) effects. Fitted models had either three or two factors (Factors 1 and 2 were combined), did or did not include method effects. CFA solutions all had no cross-loadings. Parcel solutions were constructed from the CFA models using a distributive (DIST) or homogeneous (HOMO) strategy.

b_1, b_2, b_3 = path coefficients relating Factors 1, 2 and 3 to the outcome variable. r_{12}, r_{13}, and r_{23} are correlations among Factors 1, 2 and 3. R^2 is the variance explained (which is approximately .5 for each population). CFA = confirmatory factor analysis. ESEM = Exploratory Structural Equation Model. CFI = comparative fit index. TLI = Tucker-Lewis Index. RMSEA = root mean square error approximation.

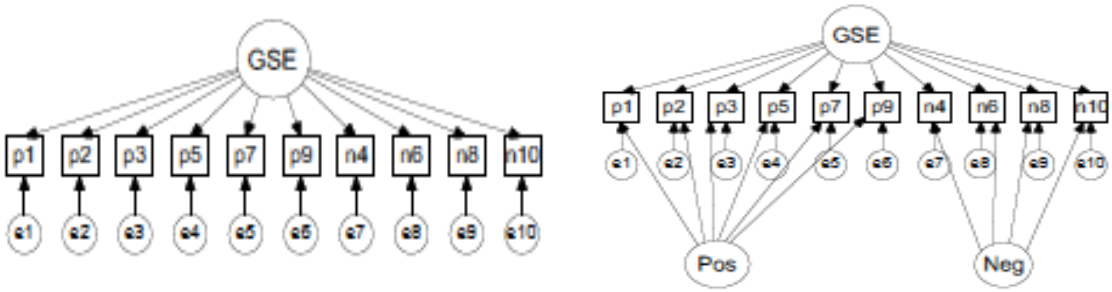


Figure 1. Two structural equation models of self-esteem for single-wave data. Model 1 posits one (unidimensional) trait (global self-esteem, GSE) with no method effects. Model 2 posits one self-esteem trait in combination with method-effect factors associated with items with positively worded item (Pos) and negatively worded items (Neg). p = positive items; n = negative items; e = error (item uniqueness). (Adapted with permission from Marsh, Scalas & Nagengast, 2010).

Supplemental Materials

NOTE: Supplemental materials presented in this section are to appear on the *Psychological Method's* External Website and hot-linked to the electronic version of the article, but are not part of the print version of the article

Appendix 1. Further Background to Study 1. The use of Parcels for Self-Esteem Responses

Appendix 2. Figure 1: 8 Models of Self-esteem Responses

Appendix 3. Table 1 Study 1: Confirmatory Factor Analysis and Tests of Invariance For Measures of Self-Esteem on Four Occasions

Appendix 4. Empirical results for factor structure at the item level for Extraversion and Neuroticism Factors (Study 2)

Appendix 5. Population generating Model used to simulate data sets in Study 3.

Appendix 6. Population generating Model used to simulate data sets in Study 4.

Appendix 1.

Further Background to Study 1. The use of Parcels for Self-Esteem Responses

Self-esteem, typically measured by some variant of the Rosenberg Self-Esteem Inventory (RSEI), is one of the most widely studied constructs in psychology. Nevertheless, there is broad agreement that a simple unidimensional factor model, consistent with the original design and typical application in applied research, does not provide an adequate explanation of RSEI responses. However, there is no clear agreement about what alternative model is most appropriate—nor even a clear rationale for how to test competing interpretations. Three alternative interpretations exist: (a) two substantively distinct self-esteem trait factors (positive and negative self-esteem), (b) one self-esteem trait factor and ephemeral method artifacts associated with positively or negatively worded items, or (c) one self-esteem trait factor and stable response-style method factors associated with item wording. Marsh, Scalas and Nagengast (2010) posited 8 alternative models (see Figure 1) and structural equation model tests based on longitudinal data (4 waves of data across 8 years with a large, representative sample of adolescents). Longitudinal models provided no support for the unidimensional model, undermined support for the 2-factor model, and clearly refuted claims that wording effects are ephemeral, but did provide good support for models positing one substantive (self-esteem) factor and two response-style method factors that are stable over time. Their longitudinal methodological approach not only resolved these longstanding issues in self-esteem research but also has broad applicability to most psychological assessments based on self-report with a mix of positively and negatively worded items. In the present investigation, we extend this research by considering the nature of conclusions for analyses based on item parcels that are based on the implicit assumption that there are no item-wording method factors.

Method: Study 1

Study 1 is based on the Marsh, Scalas and Nagergast (2010) study, where the sample, data, analyses, and alternative models are described in more detail. The Youth in Transition (YIT) data (Bachman, 2002) are based on a representative sample of 87 U.S. public high schools and approximately 25 students from each school collected on four occasions: Wave 1: early 10th grade (N = 2,213); Wave 2: late 11th grade (N = 1,886); Wave 3: late 12th grade (N = 1,799); Wave 4: 1 year after normal high school graduation (N = 1,620). Data considered here are for a 10-item self-esteem scale derived from the RSEI with six positively worded items (e.g., “I feel that I’m a person of worth, at least on an equal plane with others”) and four negatively worded items (e.g., “I feel that I can’t do anything right”). A 5-point scale ranging from almost always true (1) to never true (5) was used. For both single-wave and longitudinal CFAs, full information maximum likelihood estimator (FIML) was used to account for missing data (Enders & Bandalos, 2001; Muthén & Muthén, 1998–2006).

In the published paper we considered only two of the eight models considered by Marsh, Scalas, and Nagengast (2010). The complete set of 8 models is show in Figure 1 in these supplemental materials. Model 1 is a pure unidimensional model that posits one self-esteem factor with no method factors associated with item wording. As shown here, this model is unable to fit the data when tested at the item level. The model preferred by Marsh et al. is Model 6. It also posits one global self-esteem factor, but it includes two method factors associated with positively and negatively worded items. Models that posited only method effects associated with either positively worded items (Models 5 and 7 in Figure 1) or with negatively worded items (Models 3 and 8) were rejected by Marsh et al., as these models were not able to fit the data. They also considered a two-factor model (Model 2 with a positive self-esteem factor and a negative self-esteem factor) but rejected it because it was not able to fit the data. Marsh et al. evaluated models representing method effects as method factors (Model 6 considered here) or as correlated uniquenesses (Model 3). Although both models fitted the data well for analyses based on a single wave of data, Model 3 was shown to be unstable for both these data and in a separate simulation study. More importantly, Model 3 was unable even to test the stability of the item-wording method effects shown to exist when Model 6 was applied to the longitudinal data. On the basis of their research, they argued that Model 6 was the best model.

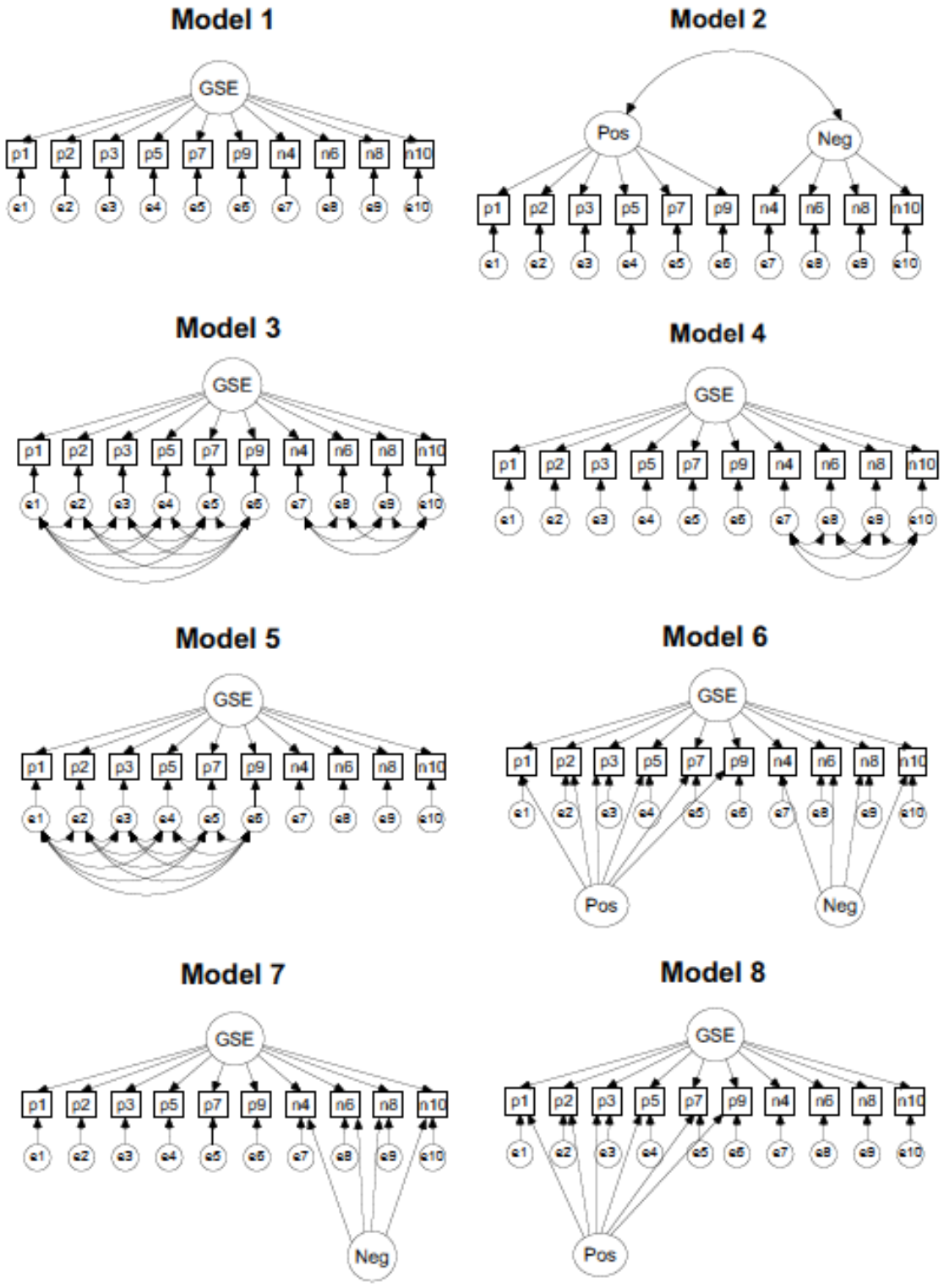
For purposes of the present investigation, we evaluated the fit of Model 1 based on item parcels as well as responses to the original 10 items. Three parceling strategies were considered. For the 5-parcel data, the first parcel was the average response to the first and sixth of the 10 items, the second parcel with the average response to the second and seventh items, and so forth. For the 3-parcel data, the first parcel was the average response to the first, fourth, seventh, and tenth items; the second parcel was the average response to the second, fifth, and eighth items; the third parcel was the average response to the second, fifth, and eighth items. These 5- and 3-parcel solutions are based on a distributed strategy in which each parcel has a mixture of positively and negatively worded items. However, because of the nature of the items, we also explored a

3-parcel homogeneous strategy in which the four negatively worded items form one parcel, and the six positively worded items form the other two parcels.

Consistent with recommendations for longitudinal panel data more generally (Marsh & Hau, 1996; Jöreskog, 1979), correlated residuals were posited a priori between matching indicators of ASC administered on different occasions (e.g., responses to the first self-esteem item administered at T1, T2, T3, and T4). Failure to control this error structure would result in positively biased estimates of stability over time and distort parameter estimates (see Marsh & Hau, 1996). It is important to emphasize that all correlated uniquenesses were posited a priori and clearly follow from previous empirical research and theory. All longitudinal models considered here included these correlated uniquenesses. However, because they are not substantively important for present purposes, they are not discussed further.

Goodness of fit. In applied CFA/SEM research, there is a predominant focus on goodness of fit indexes that are sample size independent (e.g., Hu & Bentler, 1999; Marsh, Balla, & McDonald, 1988; Marsh, Balla & Hau, 1996; Marsh, et al., 2005; Marsh, 2007) such as the Root Mean Square Error of Approximation (RMSEA), the Tucker-Lewis Index (TLI), and the Confirmatory Fit Index (CFI). Thus, for consistency with previous work, the three indices routinely provided by Mplus (Muthén & Muthén, 2010) are reported, as well as the robust $\chi 2$ test statistic and an evaluation of parameter estimates. The TLI and CFI vary along a 0-to-1 continuum and values greater than .90 and .95 typically reflect acceptable and excellent fits to the data. RMSEA values of less than .05 and .08 reflect a close fit and a reasonable fit to the data, respectively (Marsh, 2007; Marsh, Hau, & Grayson, 2005). However, we emphasize that these cut-off values only constitute rough guidelines (Marsh, Hau & Wen, 2004; also see Marsh, 2007; Marsh, et al., 1998, 2005).

Appendix 2. 8 Models of Self-esteem Responses (expanded version of Figure 1 in the published article)



Supplemental Materials Figure 1: 8 Models of Self-esteem Responses

Eight structural equation models of self-esteem for single-wave data. Model 1 = one trait factor, no correlated uniqueness; Model 2 = two trait factors: correlated positive and negative trait factors; Model 3 = one trait factor with correlated uniqueness among both positive and negative items; Model 4 = one trait factor with correlated uniqueness among negative items; Model 5 = one trait factor with correlated uniqueness among positive items; Model 6 = one trait factor plus positive and negative latent method factors; Model 7 = one trait factor plus a negative latent method factor; Model 8 = one trait factor plus a positive latent method factor; p = positive items; n = negative items; e = error. (Copied with permission from Marsh, Scalas & Nagengast, 2010.)

Appendix 3.

Study 1: Confirmatory Factor Analysis and Tests of Invariance For Measures of Self-Esteem on Four Occasions (expanded version of Table 1 in the published article)

Study 1: Confirmatory Factor Analysis and Tests of Invariance For Measures of Self-Esteem on Four Occasions

Model	χ^2	*df*	*TLI*	*CFI*	***RMSEA***
One Self-Esteem Factor, Each Wave Separately—no method effects					
1.1 wave1					
10 Items	651.57	35	.700	.767	.089
5 Parcels	14.05	5	.989	.994	.029
3 Parcels	0	0	1.000	1.000	0
3 Parcels (homo)	0	0	1.000	1.000	0
1.2 wave2					
10 Items	624.03	35	.710	.775	.095
5 Parcels	26.07	5	.974	.987	.047
3 Parcels	0	0	1.000	1.000	0
3 Parcels (homo)	0	0	1.000	1.000	0
1.3 wave3					
10 Items	539.94	35	.752	.807	.090
5 Parcels	59.10	5	.940	.97-	.078
3 Parcels	0	0	1.000	1.000	0
3 Parcels (homo)	0	0	1.000	1.000	0
1.4 wave4					
10 Items	542.04	35	.753	.808	.095
5 Parcels	16.73	5	.986	.993	.038
3 Parcels	0	0	1.000	1.000	0
3 Parcels (homo)	0	0	1.000	1.000	0
One Self-Esteem Factor, Longitudinal tests of invariance —no method effects					
1.5a Unconstrained model					
10 Items	2930.10	674	.838	.860	.039
5 Parcels	230.79	134	.987	.991	.018
3 Parcels	17.27	30	1.004	1.000	.000
3 Parcels (homo)	104.29	30	.977	.989	.033
1.5b Factor Loadings (FL)					
10 Items	2971.55	701	.843	.859	.038
5 Parcels	241.64	146	.988	.991	.017
3 Parcels	21.73	36	1.003	1.000	.000
3 Parcels (homo)	144.95	36	.971	.984	.037
1.5c FL & Variances (Var)					
10 Items	2975.17	704	.844	.859	.038
5 Parcels	243.87	149	.988	.991	.017
3 Parcels	24.68	39	1.003	1.000	.000
3 Parcels (homo)	147.49	39	.974	.985	.035
1.5d FL-Var-Uniqueness (UNQ)					
10 Items	3134.95	724	.839	.850	.039
5 Parcels	519.05	190	.961	.966	.031
3 Parcels	204.56	48	.973	.980	.038
3 Parcels (homo)	367.03	48	.937	.954	.055

Table 1 (continued)

Model	χ^2	*df*	*TLI*	*CFI*	***RMSEA***
One Self-Esteem Factor, Positive & Negative Item Method Factors; Each Wave Separately					
2.1 wave1	69.62	25	.970	.983	.028
2.2 wave2	70.62	25	.969	.983	.031
2.3 wave3	88.48	25	.956	.976	.038
2.4 wave4	78.83	25	.963	.980	.037
One Self-Esteem Factor, Positive & Negative Item Method Factors; Longitudinal Tests of Invariance					
2.5-0 Method Factors uncorrelated over time [a]	1483.37	634	.935	.947	.025
2.5a Unconstrained	916.52	622	.977	.982	.015
2.5b FFL	962.06	673	.979	.982	.014
2.5c FL & Var	1000.90	682	.977	.980	.015
2.5d FL-Var-Uniq	1161.90	702	.968	.971	.017

Note. See Figure 1 for a description of the models. Separate analyses were done for each wave separately and then longitudinal analyses across all four waves. For the longitudinal analyses, separate tests were done to evaluate the invariance of factor loadings (FL), factor variances (VAR), and uniquenesses (UNQ). The 5 Parcel and 3 Parcel models are based on parcels with a mix of positive and negatively worded item. The 3 Parcels (homo) model is based on one parcel with four negatively worded items and two parcels each containing three positively worded items. χ^2 = chi-square test statistic; df = degrees of freedom; TLI = Tucker–Lewis index; CFI = comparative fit index; RMSEA = root-mean-square error of approximation. The 5 Parcels and 3 Parcels were based on parcels with a mix of positive and negatively worded item. 3 Parcels (homo) are based on one parcel with four negatively worded items and two parcels each containing three positively worded items.

[a]In Model 2.5-0, all positive and negative item method factors were constrained to be uncorrelated. In subsequent models (2.5a-2.5d) positive time method factors were allowed to be correlated over time, as were the negative item method factors (see Table 2). However, positive and negative item method factors were constrained to be uncorrelated with each other in all models.

Appendix 4. Empirical results for factor structure at the item level for Extraversion and Neuroticism Factors (Study 2)

Empirical Parameter Estimates for Item level Big-Five Data

	CFA with Big-five Factor Loadings			ESEM Target Big-five Factor Loadings		
Items	Fac1	Fac2	Resid	Fac1	Fac2	Resid
Factor 1						
Y1	.088	.992		.076	-.030	.993
Y2	.541	.707		.493	-.128	.722
Y3	.535	.714		.561	.027	.689
Y4	.428	.817		.362	-.203	.805
Y5	.632	.601		.633	-.063	.583
Y6	.702	.507		.658	-.126	.526
Y7	.458	.790		.469	.004*	.780
Y8	.461	.788		.445	-.049	.793
Y9	.610	.627		.599	-.088	.617
Y10	.573	.672		.528	-.166	.667
Y11	.657	.569		.648	-.059	.565
Y12	.437	.809		.440	-.005*	.806
Factor 2						
Y13	.619	.617		.032	.711	.501
Y14	.360	.870		.060	.401	.843
Y15	.342	.883		-.208	.258	.874
Y16	.616	.620		-.018*	.661	.559
Y17	.415	.828		.026*	.455	.796
Y18	.337	.886		-.051	.334	.880
Y19	.449	.798		-.239	.330	.810
Y20	.625	.609		-.261	.510	.632
Y21	.547	.701		-.376	.376	.675
Y22	.170	.971		.163	.267	.916
Y23	.570	.675		-.194	.480	.703
Y24	.410	.832		-.172	.341	.836
FVar/cov						
Fac1	1.000			1.000		
Fac2	-.487	1.000		-.151	1.000	

Note. Not presented are the a priori correlated uniquenesses (see earlier discussion)
* values constrained to be approximately zero for the ESEM target rotation

Appendix 5. Population generating Model used to simulate data sets in Study 3

Population Parameters and Empirical Parameters Estimates for Item level Simulated Datasets 1 and 2

Good Simple Structure

	Population Model			CFA			ESEM Target		
Items	Fac1	Fac2	Resid	Fac1	Fac2	Resid	Fac1	Fac2	Resid
Factor 1									
Y1	.70	.00	.510	.685		.531	.700	.003*	.508
Y2	.70	.00	.510	.680		.538	.700	-.003*	.515
Y3	.60	.00	.552	.590		.652	.604	-.002	.635
Y4	.60	.00	.444	.588		.654	.604	-.006	.639
Y5	.60	.10	.444	.635		.596	.605	.098	.602
Y6	.60	.10	.444	.635		.597	.605	.097	.602
Y7	.60	.10	.316	.632		.601	.600	.098	.604
Y8	.50	.10	.465	.536		.713	.506	.093	.720
Y9	.50	.20	.465	.581		.662	.499	.200	.655
Y10	.50	.20	.330	.580		.664	.503	.196	.664
Y11	.50	.20	.330	.582		.661	.504	.200	.662
Y12	.50	.20	.330	.584		.659	.505	.202	.658
Factor 2									
Y13	.00	.70	.510		.684	.532	.002*	.702	.510
Y14	.00	.70	.510		.681	.537	-.002*	.700	.511
Y15	.00	.60	.552		.582	.661	-.002	.596	.642
Y16	.00	.60	.444		.582	.662	.000	.598	.650
Y17	.10	.60	.444		.635	.597	.097	.604	.600
Y18	.10	.60	.444		.629	.604	.101	.594	.605
Y19	.10	.60	.316		.637	.594	.102	.604	.599
Y20	.10	.50	.465		.531	.718	.100	.493	.716
Y21	.20	.50	.465		.581	.662	.202	.500	.659
Y22	.20	.50	.330		.577	.667	.199	.495	.661
Y23	.20	.50	.330		.578	.666	.210	.493	.661
Y24	.20	.50	.330		.582	.661	.201	.501	.658
FVar/covar									
Fac1	1.0			1.000			1.000		
Fac2	.25	1.0		.521	1.000		.253	1.000	

Moderate Simple Structure

	Population Model			CFA			ESEM Target		
Items	Fac1	Fac2	Resid	Fac1	Fac2	Resid	Fac1	Fac2	Resid
Factor 1									
Y1	.70	.00	.510	.617	.619		.697	.000*	.517
Y2	.70	.00	.510	.621	.614		.700	.000*	.509
Y3	.60	.10	.552	.604	.635		.602	.099	.598
Y4	.60	.20	.444	.674	.546		.605	.199	.535
Y5	.60	.20	.444	.669	.552		.598	.199	.539
Y6	.60	.20	.444	.670	.551		.599	.201	.541
Y7	.60	.30	.316	.736	.458		.597	.299	.461
Y8	.50	.30	.465	.646	.582		.500	.295	.585
Y9	.50	.30	.465	.647	.581		.496	.304	.586
Y10	.50	.40	.330	.711	.495		.497	.396	.491
Y11	.50	.40	.330	.712	.493		.499	.395	.489
Y12	.50	.40	.330	.714	.490		.505	.397	.492
Factor 2									
Y13	.00	.70	.510	.619	.616		.000*	.697	.507
Y14	.00	.70	.510	.620	.615		.000*	.701	.509
Y15	.10	.60	.552	.603	.636		.103	.600	.601
Y16	.20	.60	.444	.668	.554		.197	.600	.543
Y17	.20	.60	.444	.669	.552		.201	.596	.538
Y18	.20	.60	.444	.669	.552		.201	.597	.541
Y19	.30	.60	.316	.735	.459		.299	.597	.463
Y20	.30	.50	.465	.650	.578		.304	.504	.591
Y21	.30	.50	.465	.650	.577		.298	.502	.579
Y22	.40	.50	.330	.715	.488		.401	.500	.487
Y23	.40	.50	.330	.714	.490		.401	.498	.488
Y24	.40	.50	.330	.713	.491		.401	.497	.489
FVar/cov									
Fac1	1.0			1.000			1.000		
Fac2	.25	1.0		.836	1.000		.249	1.000	

* values constrained to be approximately zero for the ESEM target rotation

Appendix 6. Population generating Model used to simulate data sets in Study 4

	Factor Loadings					
Items	Fac1	Fac2	Fac3	Fac4	FMth	Resid
Y1	.70	.00	00			.508
Y2	.70	.00	.00			.509
Y3	.70	.20	.00			.401
Y4	.70	.20	.00			.357
Y5	.70	.20	.10			.375
Y6	.70	.20	.10			.378
Y7	.60	.20	.10	.20		.128
Y8	.60	.30	.10	.20		.220
Y9	.50	.30	.10	.20		.496
Y10	.50	.30	.10	.20		.370
Y11	.00	.70	.00			.509
Y12	.00	.70	.00			.511
Y13	.20	.70	.00			.302
Y14	.20	.70	.00			.356
Y15	.20	.70	.10			.377
Y16	.19	.70	.10			.298
Y17	.20	.60	.10	.20		.460
Y18	.30	.60	.10	.20		.393
Y19	.30	.50	.10	.20		.160
Y20	.30	.50	.10	.20		.372
Y21	.00	.00	.70			.510
Y22	.00	.00	.70			.513
Y23	.00	.00	.70			.506
Y24	.00	.00	.70			.509
Y25	.10	.10	.70			.337
Y26	.10	.10	.70			.397
Y27	.10	.10	.60			.569
Y28	.10	.10	.60			.538
Y29	.10	.10	.50			.681
Y30	.10	.10	.50			.682
Y31					.80	.357
Y32					.80	.358
Y33					.80	.358
Y34					.80	.357
Y35					.80	.360
Y36					.80	.360

Factor	Fac1	Fac2	Fac3	Fac4	FMth
Factor correlations					
Fac1	1.000				
Fac2	.60	1.000			
Fac3	.30	.30	1.00		
Fac4	.00	.00	.00	1.00	
FMth	.00	.00	.00	.00	1.00
Path Coefficients[a]					
Fac1			.556[a]		
Fac2			.000[a]		
Fac3			.306		
Fac4			---		
FMth			---		
R-Sq			.500		

Note. Population generating model used in Study 4.

a Path coefficients (β_1 & β_2) relating Factor 1 and Factor 2 (independent variables) to Factor 4 (dependent variable) varied for the three population generating models:
Population 1 ($\beta_1 = \beta_2 = 0.306$); Population 2 ($\beta_1 = 0.555$, $\beta_2 = 0$; Population 3 ($\beta_1 = 0.670$, $\beta_2 = -0.200$).