1 Introduction

Latent variable structural equation modeling (SEM; Hoyle, 2012; MacCallum & Austin, 2000; Tarka, 2018) is a common data analytic method in developmental science. An important advantage of SEM is the ability to model multi-item scales as latent constructs. Advantages of latent constructs compared to other common methods such as sums or averages of items include correcting for measurement error, making minimal psychometric assumptions, establishing factorial invariance across time and groups, evaluating model fit, and broad flexibility for confirmatory modeling (Hoyle, 2012; Lei & Wu, 2007; Tomarken & Waller, 2005; Van De Schoot et al., 2015). The benefits of SEM and of using latent constructs can be augmented with parceling, which is a pre-analytic step done before estimating latent constructs. Parceling involves making aggregates of two or more items, and then using these aggregated items as the indicators of the latent constructs (Little et al., 2002; Matsunaga, 2008). In this guide, we first provide a detailed overview of parceling and its benefits. Second, we detail methods for building parcels, particularly in the context of longitudinal models and when dealing with missing data (a ubiquitous issue in longitudinal research). Finally, we review the use of parceling in the recent developmental literature.

2 Parceling: What Is It and Why Use It?

In the measurement model of SEM and confirmatory factor analysis (CFA), latent constructs are estimated by regressing each indicator on to its respective construct (Matsunaga, 2010; Violato & Hecker, 2007). Parceling is a pre-analytic step that is done prior to estimating the latent constructs. As already mentioned, when creating a parcel (sometimes referred to as a *testlet*; Thompson & Melancon, 1996), two or more items of a construct are aggregated (i.e., averaged or summed) before being used as the modeled indicators of the latent constructs (Little et al., 2002; Matsunaga, 2008). Accordingly, any multi-item scale can be reduced to a few parcels that would then be used as the indicators to estimate latent constructs. For example, a nine-item scale could be reduced to three parcels, each consisting of three averaged items, resulting in three indicators (parcels) rather than nine indicators (items) when estimating the latent factor (see Figure 1).

Parceling is increasingly used in developmental studies but remains underutilized in a majority of studies (see Section 5). The additional step of parceling has many advantages that warrant its consideration when elaborating analytic models that use latent constructs in the SEM framework, particularly in the context of longitudinal SEM. In the next section, we discuss the advantages of parceling with a focus on cross-sectional models, but every advantage generalizes to longitudinal research, and, in fact,

Table 1 General benefits of parceling (versus using items as indicators for a latent construct)

Psychometric Benefits	Model Estimation Benefits
Higher reliability	Lower indicator to sample size ratio
Greater indicator communality	Lower likelihood of correlated residuals
Higher common-to-unique factor variance ratio	Lower likelihood of dual factor loadings
Lower likelihood of distributional violations	Reduced sources of sampling error
More, tighter, and more equal intervals	Reduced sources of parsimony error

Note. Adapted from Little et al. (2002).

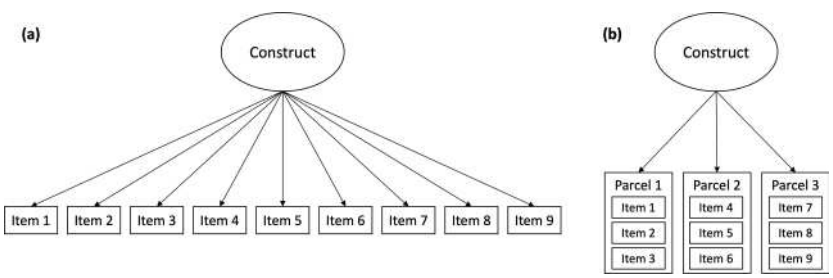

Figure 1 Latent construct with (a) nine item-level indicators and (b) three parcel-level indicators comprised of three items each. Variances, covariances, and mean structure are not included in the figure for reading ease

parcels have particular advantages for longitudinal models in their own right. For example, making complex longitudinal models more manageable, increasing the likelihood of achieving measurement invariance over time, improving model estimation efficiency (e.g., power, iteration speed, and likelihood of convergence), and others, which we discuss in more detail next.

2.1 Advantages of Parceling

Parceling can help address several problems that arise in item-level data at the psychometric level and at the modeling level (Little et al., 2002, 2013; Matsunaga, 2008). The main benefits of parceling are summarized in Table 1.

2.2 Psychometric Benefits of Parcels

2.2.1 Higher Reliability, Scale Communality, and Common-to-Unique Factor Variance Ratio

The advantages of aggregation have been discussed for decades, with the arguments for multi-item scales being closely related to the arguments for the use of parcels as indicators of latent constructs. The principle of aggregation states that aggregate scores are more likely to be representative of the construct of interest in contrast to item-level scores or single measurements (Rushton et al., 1983); moreover, aggregate scores are less biased and more statistically reliable because errors of measurement are reduced in the process. Aggregating items increases reliability, which, in turn, yields stronger associations among constructs that are theoretically related when they are modeled with aggregate scores rather than with single measurements or items (Diamantopoulos et al., 2012; Rushton et al., 1983). That is, "whenever there is the possibility of unreliability of measures, then aggregation becomes a desideratum" (Rushton et al., 1983, p. 34). The same logic, detailed later in this Element, applies to the use of parceling. Here, parceling can be considered partial aggregation of the set of items in to two or three indicators of the construct, whereas full aggregation would involve aggregating all the items into a single indicator. Full aggregation has the disadvantage, however, of not being able to separate true score information from measurement error and leads to attenuated estimates of any associations that are modeled. Multiple parcels, on the other hand, allow estimation of measurement error in the form of the residual variances of the parceled indicators.

From a traditional measurement theory perspective, item responses include several sources of variance: the desired true source of variance representing the construct of interest, as well as the undesirable variance that can come from a number of sources. This undesirable variance can arise from numerous effects such as method contamination, acquiescence response bias, social desirability, priming, and item-wording characteristics such as negatively valanced items, subordinate clauses, and common parts of speech (Little et al., 2002). By averaging items into parcels, the proportion of desirable variance related to the construct of interest is accentuated while the proportion of undesirable variance related to the effects listed above is decreased.

Both the principles of aggregation (Rushton et al., 1983) and foundational principles of psychometrics (McDonald, 1999) give a framework to understand the advantages of parceling. Figure 2 depicts a visual of the domain sampling model for selecting items to represent a construct. There are three fundamental assumptions of a domain sampling model under psychometric theory: (1) an infinite number of indicators exists in construct space; (2) a finite number of

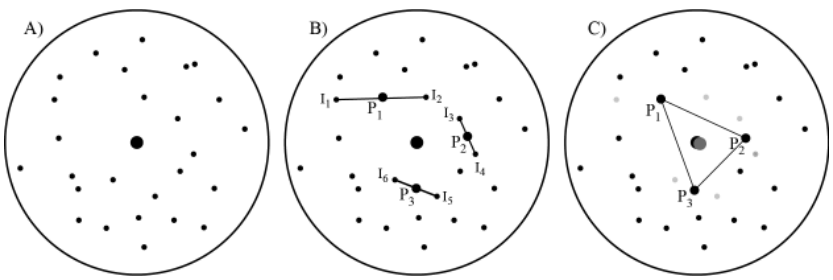

Figure 2 Domain sampling model of three parcels derived from six items.
From Little (2013, in press), reproduced with permission of the author

indicators may be selected to reflect the meaning of this construct; and (3) each indicator will have some degree of relationship with the construct's true centroid and some degree that is unrelated to the construct as well as random measurement error (Little et al., 1999).

Panel A of Figure 2 shows six items that surround the construct centroid. The outer circle represents the potential selection plane of the items for representing the construct. Panel B shows the location of each parcel when two items are averaged. Specifically, the location is the midpoint of the line linking the two items being averaged to create the new parceled indicator. Panel C shows how the three parcels triangulate around the true construct centroid. In fact, the geometric midpoint of the area defined by the three parcels is the construct centroid that the parcels measure. In the case of the parcels in Figure 2, the parcels point to the true construct centroid (see Little et al., 1999 for examples of when indictors are inaccurate in representing the construct centroid). Figure 3 shows a different view on the possible selection planes. Here, the different selection planes vary in terms of their communality (i.e., the height of the plane indicated by the vertical axis of the figure). The horizontal axis shows the width of a particular selection plane. The width of a plane is orthogonal to the height of a plane. The width of a plane is the amount of reliable variance in the indicator that is unrelated to the construct (i.e., item-specific information). The distance of a selection plane from the outer arch represents the amount of unreliability of an item selected from a given plane. In general, when items are averaged, the new parcel moves closer to the construct centroid (i.e., has less selection diversity or lower specific variance, s – which we define more thoroughly later in this section) and moves up in terms of its communality (i.e., the true score, T, of the parcel – again, we define more thoroughly later in this section). For example, items with communalities of around .4 would yield parcels with communalities around .6 and much narrower selection diversity than the items had.

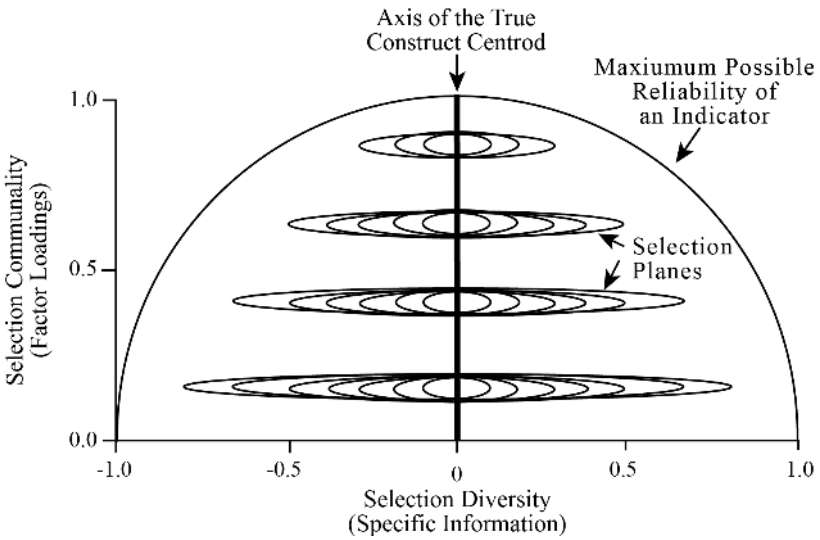

Figure 3 Side view of the domain sampling model showing various selection planes with differing levels of selection diversity (horizontal axis) and differing levels of communality (vertical axis). From Little (in press), reproduced with permission of the author

As already mentioned, modern test theory states that the score for any indicator, or variable, is composed of multiple sources of variance. Latent variable modeling seeks to decompose these sources of variance in order to increase the generalizability of a construct (Little, 2013). Modeling these sources of variance is important because some are necessary and desirable (i.e., the true score), and others are undesirable (i.e., the indictor-specific variance) as well as random noise in the indicator.

There are three general variants of test theorem tenets as shown in Equation 1:

Equation 1. Traditional test theorem variants

$$a) \quad x_i = T_i + s_i + e_i$$

$$b) \quad x = T + s + e$$

$$c) \quad x_1 = T + s_1 + e_1$$

where x_i is the score for an individual i,

i is the index referring to an individual or unit,

x without subscript is the set of scores across all individuals and units,

x_1 is the set of scores for indicator 1 of a latent construct,

T is the "true-score" variance,
s is the item- or variable-specific variance, and
e is truly random error variance (random noise).

By definition, the true-score component, *T*, the item-specific component, *s*, and the error variance, *e*, are uncorrelated with each other within a score, *X* (Bollen, 1989; Little, 2013). A key assumption of test theory is that the *s*'s and *e*'s are uncorrelated with each other across all indicators both within a scale and between scales and that they have zero means. The true-score information, *T*, across a set of items for a given construct is, by definition, the shared information, or common variance, among the items of a construct. The sum of the item-specific variance and the error variance is known as the uniqueness of the indicator, or the indicator's unique factor. The uniqueness of an indicator is broadly referred to as the indicator's residual. The assumption about the independence of the *s*'s is untested when scale averages are modeled. With latent variable modeling, however, it is possible for the item-specific variance of one indicator to have shared information with the item-specific variance from another indicator (e.g., shared wording between two items). When item-specific variances have shared information (after controlling for the common variance among the set of indicators), they can be allowed to correlate in the measurement model (i.e., correlated residuals are allowable).

Expanding upon classical test theorem, Equation 2 highlights additional sources of variance that can emerge, particularly in longitudinal studies. That is, the true score from Equation 1 can be conceptualized as having multiple components. In Equation 2, *C* represents the individual differences in the true score of the indicator that is constant across time (i.e., trait-like stability between individuals). Given that every measure is a combination of trait and method (Campbell & Fiske, 1959), *M* represents the potential of shared method-related variance in each indicator of a construct. *O* represents the change in the individual differences variance that can occur at a given occasion of measurement. Not represented in this equation is the within-person change (controlling for between-person differences) that can occur between two occasions of measurement. In longitudinal measurement, the occasion of measurement can represent age, time, or historical events. Longitudinal studies attempt to measure changes in the construct variance (*O*), without the influence of the stable construct variance (*C*), and when possible, without the influence of method variance (*M*).

When method variance is included in the measure of the construct, bias is present. When two constructs (or scales) are measured similarly, such as both

use pencil and paper method or both rely on self-report, the construct variance now represents the shared true variance as well as method variance, and the correlation between the two constructs will be inflated. This method contamination also applies if one construct is measured with a teacher observation and another with student self-report. Here, the correlation between them will be deflated by the nonshared method variance. Although the source of variance or contamination is often difficult to determine when constructs have a similar amount and type of contamination, the relative differences in correlations among the constructs can still be meaningfully interpreted, even though there will be bias in the absolute levels of the correlations.

Equation 2. A broader conceptualization of measurement theory

$$a) \quad \mathrm{I}_1 = C + M + O + S_1 + e_1$$
$$b) \quad \mathrm{I}_2 = C + M + O + S_2 + e_2$$
$$c) \quad \mathrm{I}_3 = C + M + O + S_3 + e_3$$

where I_n is the set of scores for an indicator n,
C is the stable construct variance,
M is potential common method variance,
O is the occasion of measurement variance,
S_n is the item- or variable-specific variance for indicator n,
e_n is truly random variance (random noise) for indicator n,
within an indicator, $C, M, O, S,$ and e are uncorrelated (independent),
between indicators, only S and e are uncorrelated,
sources of variance are assumed to be normally distributed, and
only, C, M, O can have nonzero means. S and e have a mean of zero.

These equations show that item values are not a perfect representation of a construct because they consist not only of the "true" score reflecting the latent construct of interest, but also of the item's specific variance and random error (as mentioned, the sum of these latter two sources of variance are referred to as the item uniqueness or the item's unique factor). Because unique variance is a component of the total variance of a total scale average (or sum), regression coefficients based on the fully aggregated scale will be underestimated (i.e., attenuated; Nunnally, 1978). Additionally, the sum of the random error and the item-specific variance reduce communality among the scale items (Gorsuch, 1988).

The three sources of variance within items are captured in a factor analytic model of the item structure. Accordingly, the true score (T) is the variance

shared among the indicators (i.e., T of each item) that defines the common factor (i.e., the modeled construct), that is, the information that we want to capture in the latent factor. Although T is shared between items, s_i and e_i are assumed to be uncorrelated with each other. As mentioned, the sum of s_i and e_i is referred to as the item's uniqueness. Theoretically, s and e should be uncorrelated with each other and have a mean of zero. Across all items in a pool of items, it is assumed that the s and e of each item are uncorrelated with all other items' s and e elements. This assumption is theoretically true for ε elements because these elements reflect the truly random error aspect of measurement, and must, by definition, be uncorrelated with all other ε and s elements. The s component of a given item, however, while uncorrelated with ε elements is unlikely to be truly uncorrelated with all other s elements. Accordingly, the common saying that "s is assumed to be uncorrelated with all other s elements" is not completely accurate. The actual assumption is that the s elements are only trivially correlated with other s's, and that they can thus be treated *as if* they were uncorrelated.

Aggregation through parceling preserves the shared variance (T) while reducing the variance related to item uniqueness, which is not shared between items. Thus, compared to items, a parcel score would have a higher proportion of true score related to the construct relative to uniqueness not related to the construct, making the score of the parcel more reliable than the score of individual items. As can be seen in Figure 4, a two-item parcel is expected to include the variance of the true score (T) plus one-fourth of each of the four sources of variance not associated with the common factor: (1) the specific variance of item 1, (2) the specific variance of item 2, (3) the random error of item 1, and (4) the random error of item 2. Thus, the variance of the parcel is expected to be equal to the shared variance of $T + \frac{1}{4}$ of the variance of s_1, s_2, ε_1, and ε_2. If three items are averaged, then each of the s and ε of the three items is reduced to one-ninth of their original magnitudes (see Little et al., 2013, for the proof of these concepts). On the one hand, since items aggregated into a parcel all measure the same construct, the T portion of variance is shared across items and captured in the parcel. On the other hand, since s and e are either uncorrelated or trivially correlated across items, both are reduced after aggregating in a parcel. Thus, by reducing the effect of item-level specific variance (s_1 and s_2) and random error (ε_1 and ε_2), parcels allow for a more accurate model of the true construct variance when compared to items.

2.2.2 Lower Likelihood of Distributional Violations

In addition to being more reliable than individual items, aggregated scores (such as parcels) tend to better approximate the distribution of the construct being

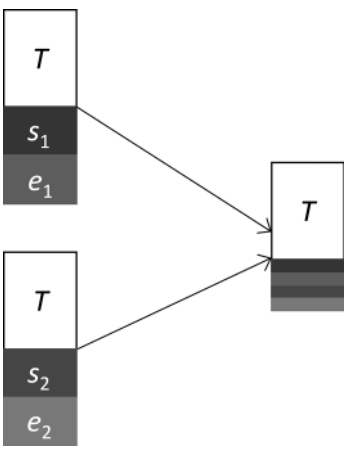

Figure 4 Variance of parcel composed of two averaged items. *T* = true score; *s* = specific variance; *e* = random error. Adapted from Rioux, Stickley et al. (2020)

measured than individual items (Boyle, 1991). In a multi-item measurement, each item usually measures only part of the construct (Rushton et al., 1983). Since individual items are expected to tap into a smaller portion of the construct than aggregated scores, their distribution is also expected to diverge more from the construct's distribution compared to aggregated scores, which better approximates the true distribution of the construct. Thus, even if a construct is theoretically normally distributed, items may show nonnormal distributions. When these items are aggregated into parcels, the resulting distribution would be more likely to approximate the normal distribution. The same principle would be true for constructs with nonnormal distributions, such as Poisson, exponential, or lognormal distributions (Galambos & Kotz, 1978; Joo et al., 2017). In line with this theoretical advantage of parcels, methodological research has shown that parcels can help remedy problems with nonnormal distributions (Bandalos, 2002; Bandalos & Finney, 2001; Hau & Marsh, 2004; Nasser & Wisenbaker, 2003; Thompson & Melancon, 1996), which can be an important advantage for analyses based on the assumption of normality. For a graphical depiction of this normalizing tendency of parcels, see Matsunaga (2008).

2.2.3 More, Tighter, and More-Equal Intervals

Parcels also have more, tighter, and more-equal intervals between scale points when compared to the scale of individual items. This advantage is particularly relevant in developmental science where many questionnaires still rely on the

Likert scale (Likert, 1932), which is essentially measuring continuous constructs with an ordinal scale. While one solution for this problem is at the measurement stage, where continuous measurements can be obtained using a visual analog scale (Rioux & Little, 2020), research suggests that although Likert-type items produce ordinal data, aggregates of Likert-type items are robust in analyses that assume the presence of interval data (Carifio & Perla, 2008). Thus, aggregating through parceling can help in having a more interval-level of measurement. Likert-scale items will only have responses at each integer along a Likert-type scale, a composite parcel based on an average will include values that fall between these integers, giving parcels a more continuous scale of measurement. By both improving the distribution of indicators (see previous section) and making ordinal measurements more continuous, parceling can thus, in many cases, allow the use of estimators based on the assumption of continuous and normally distributed variables (e.g., maximum likelihood (ML)). Direct ML estimation does not require some form of adjustment of the estimator such as using robust maximum likelihood or weighted least square mean and variance-adjusted estimators. That is, parcels reduce violations of assumptions for analyses with ML given they are more robust indicators compared to items (Lei & Shiverdecker, 2020; Li, 2016).

In terms of more-equal intervals of measurement, parcels harmonize the ordered categories to represent the response space with intervals that are more equal than the original items. For example, a four-point scale (e.g., never, seldom, often, always) represents the response space with rough intervals; that is, the distance between never and seldom is narrow in terms of the response difference, whereas the response space between seldom and often is a wider response gap. Parceling two items from such four-point scales would yield a seven-point scale with values that fall between the rough gaps of the original four-point scales.

2.3 Model Estimation Benefits of Parcels

2.3.1 Lower Indicator to Sample Size Ratio

Models with parcels have fewer indicators compared to using items. As such, the number of parameters estimated in the model is reduced. With larger models, this reduction in parameter estimates can improve model convergence and model stability (Little et al., 2013). This advantage is exemplified with a five-construct confirmatory factor analysis model, with each construct being measured with nine items (Little et al., 2013). A single time point and single group analysis of this model would have 3,825 degrees of freedom and 270 parameter estimates (not including potential correlated residuals and cross-loadings), which would be