don't use the mean and standard deviation within a given time point, because you lose any information about change in means and variances). If you have multiple groups and longitudinal data, you would calculate the grand mean across time and groups, as well as the grand standard deviation across time and groups. With this method you can leave the scores in a grand *z*-score metric (again, however, the mean and standard deviation used must be the overall mean and standard deviation so you don't lose change information). You can also add a grand constant and multiply by a selected value to put the indicators on a so-called *t*-score metric (e.g., mean of 100 and a standard deviation of 15, or a mean of 10 and a standard deviation of 2, or a mean of 50 and a standard deviation of 10—there is no official definition of the mean and standard deviation of a *t*-score as far as I know). The choice of the rescaling method here does not "normalize" the data, nor does it change the shape of the distribution or the strength of an association between any of the variables. Rescaling simply provides a metric that makes the estimated values more interpretable (and can help with convergence problems). Of course, rescaling variables when they are already in meaningful and interpretable metrics may not be desired unless you are experiencing convergence problems.

The final recoding issue is reverse-coding of variables. I generally recommend that all indicators be coded so that a high score on each indicator has the same interpretation (e.g., more *X* or less *Y*). For example, if I have one indicator of anxiety that is worded as something like "I don't get nervous" with a Likert scale of *disagree* to *agree*, a high score would mean less anxious. If the other indicators are worded such that a high score means more anxious ("I am easily aroused"), I would reverse-code the "I don't get nervous" item. If the Likert scale is a 1–7 scale, I can simply create a new variable that is 8 minus the values of the original variable (8 – 7 = 1 and 8 – 1 = 7). It is also a good habit to name the latent construct by what the higher values mean. For example, if all my indicators are coded such that a high score means less anxiety, I would call the construct "lack of anxiety" or "nonanxiousness." If the indicators are coded such that a high score means more anxiety, I would label the construct "anxiety" or "anxiousness." In a similar vein, when creating dummy codes to represent gender or ethnic categories, I recommend labeling the variable with the meaning of a high score. For example, if gender is coded 0 = female and 1 = male, then I recommend calling the variable "male." Then you don't have to ask "Which gender is coded as 1?"

PARCELING

Parceling refers to taking two or more items and packaging them together (i.e., averaging them), much like a parcel you would take to the post office. The parcel (instead of the original items) is then used as the manifest indicator of the latent construct. Parceling is a premodeling step that is done before the data are fed into the SEM software. When packaging items to form a parcel (or a scale score for that matter), I

strongly recommend averaging the items as opposed to summing them. If you take sums and the number of items going into a parcel differs, the parcels will have different metrics, giving materially different means and variances. If you average the items, the parcels will have roughly similar metrics with similar (and comparable) means and variances. Moreover, the scores on the parcels will reflect the actual scale that was used to record the item-level information. The original scale is usually meaningful, and it provides a point of reference for interpreting the mean levels and variances.

Parcels have a number of statistical and analytical advantages over item-level analyses. Parcels also pose some challenges; the whole idea still engenders some debate in the methodology community. The debate can generally be traced to two opposing philosophical camps. The con arguments stem from the strict empiricist traditions of classical statistics. As my colleagues and I put it (Little, Cunningham, Shahar, & Widaman, 2002), the strict empiricist viewpoint suggests:

> Parceling is akin to cheating because modeled data should be as close to the response of the individual as possible in order to avoid the potential imposition, or arbitrary manufacturing of a false structure. (p. 152)

We went on to describe the alternative, more pragmatic viewpoint:

> Given that measurement is a strict, rule-bound system that is defined, followed, and reported by the investigator, the level of aggregation used to represent the measurement process is a matter of choice and justification on the part of the investigator. (pp. 152–153)

If your philosophical view is aligned with the former, you can skip to the next section. If you are undecided on the issue, I recommend you read Little et al. (2002) and Little, Rhemtulla, Gibson, and Schoemann (in press). Both of these papers describe the con arguments and provide reasoned arguments as to when the con arguments are not applicable. My most important admonition here is to be thoughtful when you create parcels. If you are thoughtful in parcel creation, parcels have many advantages and avoid the potential pitfalls that the con arguments highlight.

The motivation to create and use parcels is that they possess a number of advantages. These advantages can be summarized into two classes: their fundamental psychometric characteristics and their behavior when estimating a model. These advantages are summarized in Table 1.6.

In terms of the psychometric advantages of parcels over items, the first three that are listed in Table 1.6 (higher reliability, greater communality, higher ratio of common-to-unique factor variance) are saying essentially the same thing. Per the principles of aggregation, parcels will have greater reliability than the items that are used to create them. As a result of having greater reliability, parcels will have more true-score variance than items, which will also make the factor loadings stronger

TABLE 1.6. Key advantages of parcels versus items

Psychometric characteristics
Parcels (as opposed to the items) have . . .
• Higher reliability
• Greater communality
• Higher ratio of common-to-unique factor variance
• Lower likelihood of distributional violations
• More, tighter, and more-equal intervals
Model estimation and fit characteristics
Models with parcels (as opposed to the items) have . . .
• Fewer parameter estimates
• Lower indicator-to-subject ratio
• Lower likelihood of correlated residuals and dual factor loadings
• Reduced sources of sampling error

Note. These advantages pertain to the smaller set of parcels that are made up of a larger set of items. The advantages accrue based on the principles of aggregation and the law of large numbers.

(increased communality) and the unique factors smaller. As a result, the ratio of common-to-unique factor variance will be higher. All three of these related features of parcels are a good thing when it comes to fitting an SEM model. Regarding the advantages of the distributional properties, parcels are more likely to be normally distributed than are items.

When quantitative specialists discuss the conditions under which parcels can be used, many would agree that parcels are not problematic when the indicators are at least congeneric in nature (i.e., are truly indicators of the construct) and the construct is unidimensional "in the population." *In the population* means that if we had access to the whole population we would be able to tell which indicators go with which constructs and what each construct is truly made of—the kind of thing only an omniscient being or higher deity would know. We, as mere mortals, never know what exactly things are like *in the population*; we can, however, infer things about the population from the sample. Unfortunately, sampling variability comes into play and can muddy the waters a bit. Sampling variability is the inherent variability around the true population values for a given parameter estimate of any statistical model that is produced when you draw repeated samples from the population (or, more precisely, any given sample may deviate from the population values to some degree). On average, any given randomly drawn sample will provide estimates of the population parameters that are equal to the true population values, but there will be variability in these estimates from sample to sample. The larger the sample size,

the lower the sampling variability; the more homogeneous the population, the lower the sampling variability.

In Figure 1.3, I have presented a geometric representation of how parcels can work. In the circle denoted A, I have a "universe" of possible indicators for a construct. The construct's true center is depicted by the large dot in the center of the circle. The small dots are possible indicators that can be selected to represent the construct. Moving to circle B, I have selected six possible indicators. Numbering clockwise starting in the upper left of the area, I have assigned indicator 1 (I_1) and Indicator 2 (I_2) to be in parcel 1 (P_1). I have assigned I_3 and I_4 to be in P_2. I_5 and I_6 are assigned to P_3. In this geometric representation, dots that are closer to the centroid will have much larger loadings on the construct than will dots that are farther away. Any dots that are closer to each other will have higher correlations with each other than they will with other dots that are farther away. Given these simple geometric properties, we can see that I_2 and I_3 would be more highly correlated with each other than they would be with the actual construct they are supposed to measure. In fact, an analysis of the item-level data would result in a correlated residual between I_2 and I_3. In addition, because I_1 is quite far away from the centroid of the construct, it is likely to have a secondary loading on another construct.

If you take the average of the indicators that were assigned to each parcel, you would create a new parcel-level indicator that is at the midpoint of the line connecting the two indicators. The location of the larger dots labeled P_1, P_2, and P_3 in circle B are the locations of the parcel-level indicators that would result from averaging the corresponding item-level indicators. In circle C of Figure 1.3, I have connected the three parcels with lines to depict the triangulation on the centroid that the three

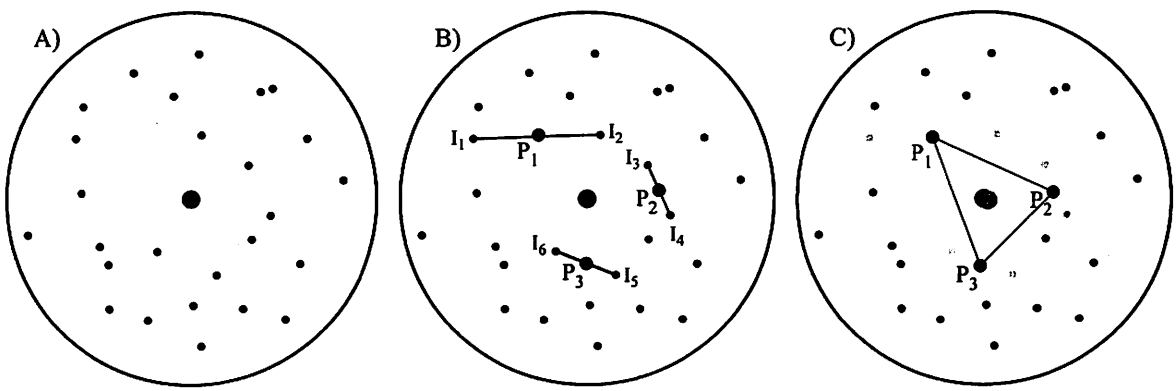

FIGURE 1.3. A geometric representation of how parceling works. Each circle represents the domain of possible indicators of a construct. The construct's "true" centroid is the larger dot in the center of each circle. The average of any two variables would be the midpoint of a straight line, as depicted in B; the average of three or more indicators would be the geometric center of the area that they encompass. The latent construct that would be indicated by the parcels is the center dot, in gray, that nearly overlaps the true centroid, as in C.

parcels would now measure. Here, the geometric midpoint of the triangle created by the three parcels is the estimated factor centroid, which, as can be seen, is close to the "true" centroid.

The parceling example that I have depicted in Figure 1.3 is not necessarily the optimal set of parcels that could have been created from the chosen indicators. Because we don't know the true location of a construct's centroid, we can only try to approximate an optimal solution given the items that were selected to represent the construct in the first place. I generally have found good results using a balancing approach whereby I assign the item with the highest item–scale correlation to be paired with the item that has the lowest item–scale correlation. I then select the next highest and next lowest items to be in the second parcel and repeatedly assign items in this manner until all items have been assigned to a parcel. In longitudinal models the selection of items to parcels needs to be the same at each time point. Here, I have used the average loading across the time points to create the rankings of the items. I generally do not recommend random parceling. Random parceling would be OK under conditions where there are lots of items and the items are all equally good (i.e., the assumption of tau-equivalent or parallel indicators), but usually we have indicators that are only congeneric (i.e., some items are better than other items).

Indicators of a construct are congeneric when they are truly indicators of the construct (in the population) but can vary in terms of the size of their loadings and intercepts. Most factor models and SEM models assume that indicators are congeneric, which is the least restrictive of assumptions about indicators. In my experience, most indicators in the social and behavioral sciences are only congeneric. As mentioned, the assumption of tau-equivalent or parallel indicators is more restrictive. Indicators of a construct are tau equivalent when their loadings are all at about the same level (again, in the population), but they can vary in terms of their intercepts. The strictest assumption about a set of indicators is that they are parallel in the population. Here, the loadings and the intercepts are all essentially equal. Indicators in the ability-testing domain can sometimes achieve this level of precision across items.

One key assumption in all this is that the statistical model being fit to the sample is the correct model for the data. In other words, we assume that the model would be true *in the population.* As MacCallum and Austin (2000; see also Box, 1979) point out, however, "all models are wrong to some degree, even in the population, and the best one can hope for is to identify a parsimonious, substantively meaningful model that fits observed data adequately well" (p. 218). So there is always some variability due to the fact that my model is going to be wrong to some degree, anyway (hopefully not too much . . .).

Parceling reduces both the sampling variability of the selected sample and the amount of incorrectness of my model in the population. The benefits of reducing the likelihood of correlated residuals and dual-factor loadings are both aspects of how parcels reduce sampling variability (a dual loading or correlated residual could be

just a sampling fluctuation) or population misfit (a dual loading or correlated residual could be true of the item-level data in the population, but it is no longer true of the parcel-level data in the population). In other words, there is a true model in the population for item-level data that probably has some "true" correlated residuals and dual loading that we might be tempted to not estimate because the misfit is mistakenly attributed to sampling variability. A true model also exists in the population for the parcel-level data, but this model is less likely to have those "true" correlated residuals and dual-factor loadings. The principle of aggregating the true-score variances of items while reducing their uniquenesses is the reason that parcels have this quality.

I have found parcels to be extremely useful and effective in nearly all circumstances, even when the construct I'm working with is multidimensional (see Little et al., in press). There are times that I won't parcel items, such as when my empirical question is about the behavior of the items across two or more samples or two or more time points. On the other hand, when the substantive questions are about the constructs and the possible differences in those constructs (e.g., differences across time or across groups), then creating parcels to use as indicators of the constructs is justified. Transparency and ethical behavior are important when using parcels. In terms of transparency, a researcher needs to be clear and honest about what he or she did when creating the parcels and why the parcels were used. The ethical part comes in when one plays around with problematic items until a particular parceling scheme works out such that a preferred outcome materializes that would not have otherwise. I trust that you are not the unscrupulous researcher that ethicists worry so much about.

WHAT CHANGES AND HOW?

Many if not most developmental change trends, if measured across the lifespan, will have a nonlinear function. Most developmental studies, however, are unable to cover the entire lifespan. When such studies cover only a limited amount of the lifespan, the resulting estimate of the trend line will often be closely approximated as a linear trend. In Figure 1.4, I have depicted two stylized and hypothetical examples of lifespan trends (these could also be other trends of change, such as trends in response to therapy or trends in hormones in response to a stressor). In panel A, the true nonlinear trend is depicted. In panel B, the observed trend lines from three different studies covering three different segments of the lifespan are pieced together. Each individual study was able to detect and therefore depict the trend line only as a linear trend. When pieced together, however, the trend lines do a rather good job of tracing the true nonlinear trajectory. The fact that nonlinear trends can be represented reasonably well with segmented linear trends should not be considered a damnation of one's theory. To detect a nonlinear trend, the important part to measure is, in fact, at the extremes, well beyond a bend (contrary to some who argue to oversample at