METHODOLOGICAL ISSUES IN MEASURING THE DEVELOPMENT OF CHARACTER

Noel A. Card
University of Connecticut

In this article I provide an overview of the methodological issues involved in measuring constructs relevant to character development and education. I begin with a nontechnical overview of the 3 fundamental psychometric properties of measurement: reliability, validity, and equivalence. Developing and evaluating measures to ensure evidence of all 3 psychometric properties has substantial impact on the quality of character development and education research. I then offer specific suggestions for using prior psychometric evidence in planning studies, for adequately reporting psychometric properties of measures used in studies, for conducting studies with a primary focus on evaluating psychometric properties, and for using meta-analysis as a methodological tool for synthesizing existing evidence of psychometric properties of scales.

Studying the development of character requires adequate measure of the constructs considered. The constructs of potential interest in studying character include character strengths and virtues, positive behaviors, supportive relationships with others, or adaptive perceptions and ways of coping with one's environment. When studying character development and education, an additional consideration is how these constructs change across time, and measures used must be sensitive to potential changes across development or educational intervention. In this article, I do not explicitly focus on any one aspect of character, character development, or character education. Instead, the broader goal of this article is to describe some more general considerations of what constitutes good measurement of a construct (i.e., any aspect of character or character development), and practices that I believe would advance the scientific study of character development and education.

I will begin by describing three fundamental psychometric properties of any measure (i.e., attributes of a measurement process). I will then describe the reasons why using measures with these properties is beneficial to studying character development and education. In the third section of this article, I will briefly describe two very different situations of research evidence of psychometric properties, illustrated by my ongoing synthesis of mea-

• **Correspondence concerning this article should be addressed to:** Noel A. Card, noel.card@uconn.edu

Journal of Character Education, Volume 13(2), 2017, pp. 29–45 ISSN 1543-1223

sures of gratitude and humility. In the fourth, and longest section, I will offer suggestions for primary research and research synthesis to advance understanding of the psychometric properties of measures relevant to character development and education.

FUNDAMENTAL PSYCHOMETRIC PROPERTIES OF GOOD MEASURES

In order to establish a foundation for considering measurement of character, it is useful to describe three fundamental psychometric properties on which measures can be evaluated: reliability, validity, and equivalence. This section offers a brief, conceptual overview of these three properties (for more extensive coverage, see Card, 2013; Little, Lindenberger, & Nesselroade, 1999; McDonald, 1999; Nunnally, 1978). An understanding of these properties is useful when evaluating empirical work measuring character, selecting or adapting existing measures into one's research, and when developing a new measure of character (a full description of the process of scale development is outside the scope of the current paper; for an accessible introduction see DeVellis, 2003). This overview will be built upon in subsequent sections to apply specifically to measuring character.

Reliability. The psychometric property of reliability refers to the repeatability across multiple measures of a construct. There are different types of reliability, including internal consistency reliability, test-retest reliability, and interinformant reliability.

Internal consistency reliability assesses the degree to which multiple items on a scale are overlapping (i.e., covary), and therefore measuring a common source of between-person variability (i.e., one typically only computes internal consistency with items assumed to represent a single construct). Internal consistency reliability is typically quantified with Cronbach's alpha, α ($\alpha = \frac{j\bar{r}}{(1 + (j-1)\bar{r})}$; where j is the number of items and is the average correlation among items). Despite the common use of Cronbach's α, it is not always the best measure of internal consistency. Cronbach's α is computed under the assumption of parallel items, which is a highly restrictive assumption imposing equal variances across items and equal correlations of items with the total scale score (e.g., McDonald, 1999). This assumption implies that every item on a scale measures the desired construct to the same degree, and that each item also assesses undesired constructs (e.g., social desirability) equivalently and contains the same amount of item-specific variance. These assumptions could be tested through factor analyses, but most researchers do not routinely test these underlying assumptions of Cronbach's α. Other indexes, such as McDonald's omega (ω; see McDonald, 1999) might better represent the underlying assumptions of measures.

Test-retest reliability is the overlap of individuals' scores measured at two different occasions. It is typically quantified as a correlation between scores at Time 1 with scores at Time 2. An important assumption of test-retest reliability that is rarely considered is that the time span between measurement occasions must be short enough that the underlying construct does not change. If this assumption is true, then the lack of correlation (i.e., correlations less than 1.0) indexes the unreliability of the measure. However, if this assumption is false, then the lack of correlation is due to both the instability of the construct and the unreliability of the measure, and it is not possible to separate these two sources of imperfect correlation. A challenge of making assumptions of the stability of character strengths, and likely many other constructs relevant for the study of many aspects of character, is that there is little research on their stability across time. Therefore, there is inadequate empirical evidence on which to base decisions of appropriate time spans over which test-retest reliability can be tested.

Interinformant reliability refers to the overlap in individual's scores between two separate reporters. For example, two different teachers might be asked to report a student's behavior,

and the correlation between these two reports quantifies the interinformant reliability. Interinformant reliability might also refer to overlap between different types of informants (e.g., teacher and peer reports of student behaviors) or informants across different contexts (e.g., classroom teacher and after-school program facilitator). Just as test-retest reliability was predicated on the assumption that the construct is stable across the time span of the multiple measures, interinformant reliability is based on the assumption that the construct is stable across the opportunities for the multiple informants to observe the behavior, and how the informants would interpret the behavior. For instance, it would only be reasonable to assess reliability between a classroom teacher and an after-school program facilitator if the student exhibited the same behaviors for both informants to observe and both informants made the same interpretations of the behavior. However, if the behaviors differ across these contexts (due to e.g., different peers, different environmental demands, different standards for behaviors, different fatigue throughout the day), then a low correlation between reporters represents cross-contextual instability of the behaviors as well as potential interrater unreliability. Thus, quantification of interrater reliability as the correlation between reporters implicitly assumes that the construct is stable across context and time, and to the extent that the construct is unstable the interinformant reliability will be attenuated. For aspects of character for which the cross-contextual and longitudinal stability is unknown, it is unclear to what extent any interinformant reliability estimates quantify instability of the measure versus construct.

Although all three types of reliability are important, most research focuses on internal consistency reliability. There are two reasons for this focus. First, internal consistency reliability is a prerequisite for other types of reliabilities if researchers conduct manifest variable analyses (i.e., as opposed to analyses using latent variables, such as confirmatory factor analysis and structural equation modeling). The reasons for this is that a lack of internal consistency will attenuate (i.e., reduce magnitudes of) correlations that the composite variable (i.e., average of multiple items) can have over time or reporter. However, latent variable analysis, which can correct for internal consistency unreliability by considering only the overlapping variance among items, could allow researchers to evaluate other forms of reliability even if internal consistency reliability is suboptimal. A second reason for this focus on internal consistency reliability is likely pragmatic; it is far easier to administer a multi-item scale to participants than it is to administer a scale across multiple reporters or across multiple occasions.

Before concluding this section, two cautions for research merit mention. First, it is important to remember that all reliability estimates are population parameters that are estimated from a sample. Reliability is not a property of a scale, but instead it is a property of the scale when applied to individuals of a particular age, in a particular context, in a particular intervention condition (etc.). Because it is an estimate, it can also vary between different samples from the same population. Therefore, it is important to estimate, interpret, and report internal consistency in every study, and not assume that evidence of reliability from previous research applies to other situations.

A second caution is against overemphasizing reliability, when more emphases should be given to validity and measurement equivalence. The commonly repeated adage that one cannot have validity without reliability is true only when one uses manifest variable analyses (because the unreliability attenuates any possible association of a manifest variable with other variables). However, unreliability is relatively easily corrected through latent variables analyses (i.e., confirmatory factor analysis and structural equation modeling; see Little et al., 1999). It is possible that efforts to establish internally consistent measures of character come at the expense of reducing validity.

Validity. The psychometric property of validity refers to the extent to which a mea-

surement instrument assesses what the researcher intended to assess. For example, to what extent does a measure of prosocial behavior actually assess individual differences in the frequency of prosocial behavior, versus assessing either a limited subdomain (e.g., helping the teacher) or irrelevant constructs outside of the operational definition of the construct. There are many potential irrelevant constructs (e.g., social desirability, peer status, reputation, academic achievement), many of which may be more or less relevant for different measures used in the study of character development and education.

One way to conceptualize validity is within the domain representation framework (Little et al., 1999; Nunnally, 1978), shown in Figure 1. In this framework, the construct of interest to have a "centroid," or exemplary definition along which individual differences in the construct are defined. Around this centroid, there exists broader domain space containing all characteristics or behaviors that might be considered as part of the operational definition of the construct; any characteristic or behavior that falls within this circle is considered part of the construct, whereas anything falling outside of this circle is not part of the construct. Many constructs in the study of character development and education (indeed, many social sciences) have fuzzy boundaries. This lack of clarity is not in itself problematic, but it can make it challenging in reaching consensus about whether an item is a useful indicator of a construct.

A valid measure of a construct is one in which the center point of the multiple items (or

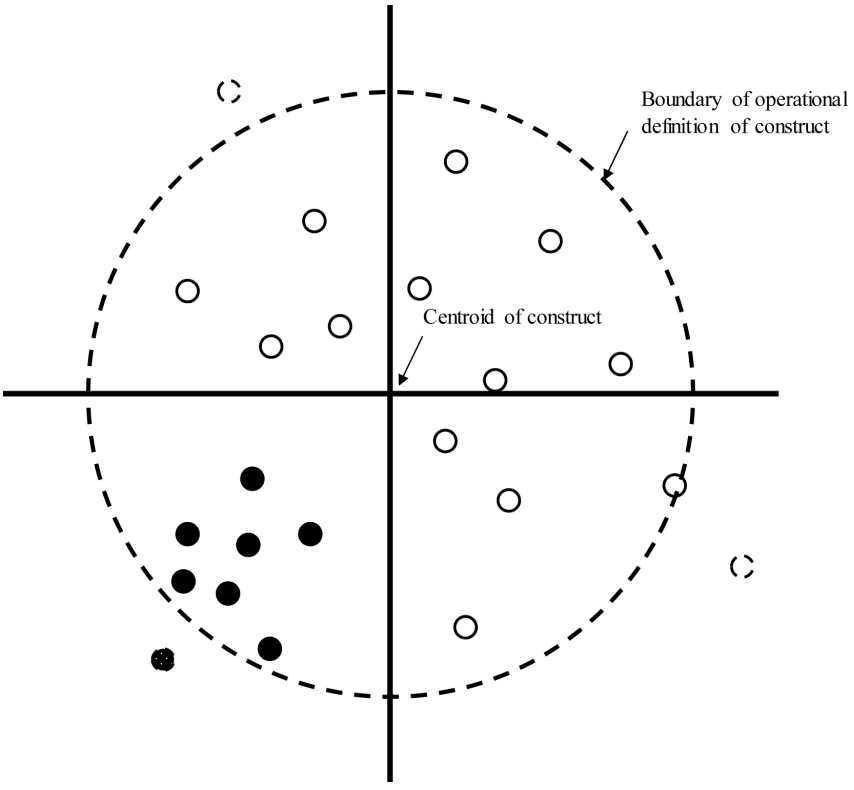

FIGURE 1
Domain Representation Framework to Illustrate Validity

multiple measurement occasions, multiple informants) is as close to the construct centroid as possible. An example of multiple items on a scale are denoted in Figure 1 with circles. If one forms a scale from these multiple items, the validity of the scale refers to the extent that their midpoint is close to the construct centroid: The closer the midpoint, the higher the validity, whereas the further the midpoint from the centroid, the worse the validity. Figure 1 also displays reliability information. Items (circles) that lie close to the centroid have a higher item-total correlation than those that lie further away from the centroid. A collection of multiple items that are close to one another will have high internal consistency reliability, whereas a collection of items that are more separated will have lower internal consistency reliability.

Figure 1 displays three indicators that fall slightly outside the boundary of the operational definition of the construct. These might be items that do not fall within the operational definition of the researcher (e.g., the researcher is using a measure developed by someone with a different theoretical or conceptual understanding of the construct). Alternatively, these might be items that contain a large amount of variance from other sources, such as a similar construct, an unintended construct (favorable reporting bias), or idiosyncratic wording of the item (e.g., poor translations of the scale across languages). It is interesting to note that these three items outside the circle of Figure 1 result in greater spread among the items (lower internal consistency reliability) but do not impact the location of the midpoint of the items (i.e., the validity).

Figure 1 also shows circles that are filled and circles that are unfilled, and is intended to depict a problematic practice that results in overemphasis on internal consistency reliability. Imagine that a researcher develops a new measure of a construct that consists of 23 items depicted as the 23 circles in Figure 1. If the first analysis performed examines item-total correlations, the researcher might decide to remove the 3 items outside the circle from the scale due to their low item-total correlations. This decision would be reasonable, especially if accompanied by a conceptual consideration of these items (e.g., a post hoc realization that an item might actually assess a different construct, be too difficult for the age of participants, etc.). If the researcher then used the 20 remaining items, the scale midpoint would align well with the construct centroid, and the measure would be valid. However, if the researcher overemphasizes internal consistency reliability, then there might be the temptation to begin removing other items because doing so increases the value of Cronbach's α. An extreme, but not necessarily implausible, example of this practice is denoted in Figure 1 with the unfilled circles representing items that were removed, and the filled circles representing items that were retained in the final scale. The filled circles (items of the final scale) are very close together, and would therefore provide a very high estimate of internal constancy reliability. In fact, this hypothetical researcher might even bring back the item outside the operational definition of the construct if doing so further increases Cronbach's α (which is a function of both average interitem correlation and the number of items of a scale). The end result is that the midpoint of these items is very clearly far from the construct centroid; so the researcher in this hypothetical scenario has sacrificed validity for the sake of reliability.

Unlike reliability, there is usually no single index to quantify validity. In principle, validity is quantified as the correlation between scores on the measure and the construct it is intended to measure, a type of validity termed criterion validity. An example of criterion validity might be a study that compares an easy, inexpensive, and/or noninvasive measure with a more challenging, expensive, or invasive measure, such as the correlation of an Alzheimer's Dementia screening instrument with later postmortem analysis. However, researchers of character development and education typically does not have values on the construct itself, therefore precluding direct assessment of criterion validity. Instead, validity is determined by

examining multiple correlations of the measure with other variables, and assessing how well these correlations match expected associations of the construct with other constructs. Evidence for this type of validity, often called construct validity, comes from both the existence of correlations between measures of the same or similar constructs (i.e., convergent validity) and from the absence of substantial correlations of measures of constructs expected to be uncorrelated or weakly correlated (i.e., discriminant validity; see Campbell & Fiske, 1959). For example, it might be expectable that a new measure of gratitude is correlated with existing measures of gratitude and with measures of spirituality, and finding of substantial positive correlations of the new measure with existing measures would be evidence of convergent validity. Continuing this example, it might be hoped that this new measure of gratitude does not correlate with a scale measuring favorable self-presentation biases, and the research would evaluate the correlations between the gratitude measure and measures of presentation biases with the expectation of extremely small and/or nonsignificant correlations.

As described further below, many studies of character strengths report internal consistency reliability, but few report validity correlations. I anticipate that this situation is similar in other areas of studying character development and education, and the result is that many efforts in studying character have pursued reliability to the relative neglect of validity. One reason has been the reliance on manifest variable analyses, in which variables must be reliable in order to have sizable and consistent relations to other variables (because unreliability tends to attenuate manifest variable correlations toward zero). However, when researchers develop items for a measure based on an operational definition but then remove items in order to achieve acceptable reliability, it is likely that the refined collection of items, while more homogeneous (i.e., reliable) no longer center around the construct centroid, but instead target a narrow region of the domain space. It is worth considering whether the highly reliable measures of various aspects of character adequately assess the breadth of constructs in which we are interested.

Measurement Equivalence. In addition to reliability and validity, a third psychometric property that must be considered is measurement equivalence. Measurement equivalence refers to a measurement instrument performing in the same way across contexts, time, and/or groups. The concept is referred to by different terms, including factorial equivalence, measurement invariance, and differential item functioning (which is the absence of measurement equivalence).

There are three levels of measurement invariance that are typically of interest. The first level of invariance is often called "configural invariance." In a model assessing this level of invariance, the same items load on the same constructs across two or more situations, whether these multiple situations are different ages, different contexts, of different intervention conditions. This initial level of measurement invariance is typically the first model evaluated by the researcher, and becomes the basis for building more restrictive and rigorous models. The second level of measurement invariance, often called "weak invariance" (AKA metric invariance, loading invariance; Card & Little, 2006; Little, 1997), adds restrictions to the initial configural invariance model that the factor loadings are equated across conditions. Conceptually, this model aligns the measurement centers on the same point in domain representative space across two or more situations. If measures are found to exhibit this level of invariance across situations, then it is meaningful to compare latent variances and associations (correlations, regression paths) across situations. The third level of measurement invariance is still more rigorous, and is often called "strong invariance" (AKA scalar invariance, intercept invariance). Strong invariance builds upon the weak invariance model by also imposing relative equalities of the indicator intercepts across situations. Imposing these equalities evaluates

whether items with relatively high means in one situation are the same items with relatively high means in another, even if the overall (construct level) mean changes across situations. If measures exhibit this level of invariance across situations, it is also possible to compare latent means across situations. When measures are found nonequivalent, this is because one or more of the items functions differently across situations because some items change in meaning across age or contexts, or an intervention impacts specific aspects of some items differently than the underlying construct.

Measurement equivalence can be empirically evaluated rather than assumed. Analyses must be performed on measures with multiple items, and at the item level. The exact method of assessing measurement equivalence depends on the level of measurement of items. Item response theory analyses are applied when items are dichotomous or when there are few response options. Confirmatory factor analysis is used when items are continuous or contain enough ordinal responses to warrant treatment as continuous (a rough recommendation is that five or more response levels are sufficient in many cases; see Rhemtula, Brosseau-Liard, & Savalei, 2012). Rather than describing the details of these analyses here, readers can refer to any of several accessible introductions for more information (e.g., Card & Little, 2006; Cheung & Rensvold, 2002; Little, 1997; Millsap & Cham, 2012).

RELEVANCE OF HIGH QUALITY MEASUREMENT TO THE STUDY OF CHARACTER DEVELOPMENT AND EDUCATION

Careful consideration of these three psychometric properties (i.e., reliability, validity, equivalence) is relevant to virtually all social sciences. These considerations are no less relevant to the study of character development and education, and this section identifies aspects of studying character that pose special challenges to assessing and establishing the psychometric properties of measures in this area.

The first challenge is that the constructs considered are highly diverse, often have multiple definitions, and in many cases have extremely fuzzy boundaries of operational definitions. Figure 1 can again be used to illustrate two implications of this statement. First, if different researchers define a character strength or relevant behavior differently, this means that they are attempting to measure different construct centroids. If they actually do measure different centroids, then development trends, associations with antecedents and consequences, and responses to different interventions will differ depending on the construct studied, making it harder to synthesize across studies to draw generalized conclusions. A second implication has to do with the fuzzy boundaries of definitions: If the outer edge of a construct (i.e., the outer circle in Figure 1) are not well-defined, then there is no definitive criterion to guide whether an item assesses a construct or not. As such, researchers are likely to make empirically driven decisions, and given the tendency to prioritize reliability described above, might be tempted to prune items to maximize reliability but at the expense of validity. These two problems suggest that clarity in operational definitions, and selecting or developing measures that correspond to those definitions, are critical aspects of adequately measuring character development and education.

The second challenge involves the diversity of populations and contexts in which character development and education is studied. On the one hand, this diversity is a strength of the field in that it advances our understanding of the generalizability or differences in effects. On the other hand, we should keep in mind the earlier caution involving reliability that psychometric properties are estimates of how a measure functions in a particular population, in a particular setting, and under particular methodological conditions. Therefore, a challenge of much of character development research is to assess constructs equivalently

across development and context. An item or method of assessing prosocial behavior at one age might capture little variance (because no one enacts the behavior) or only irrelevant variance (e.g., social desirability in reporting) at another age. Similarly, some items may perform differently in one context than another; for example, an item assessing curiosity within a classroom versus playground versus the home could plausibly differ. Capturing the underlying constructs across development and contexts therefore requires explicit consideration and empirical evaluation of measurement equivalence.

A third challenge in studying character development and education is that the field consists of both naturalistic studies quantifying aspects of character, as well as intervention studies attempting to change those constructs. Again, although this is an overall strength of the field, it does require that measurement equivalence also be assessed across intervention conditions (here I will focus on between-group comparisons, though this logic applies to other, less commonly used, approaches such as pre- and postregression discontinuity, and other ways that intervention effects are studied). It is necessary to evaluate measurement equivalence across intervention conditions to ensure that interventions change the underlying construct, rather than merely changing the measurement properties. Recognition that interventions can change the construct or the measurement of the construct leads to several possibilities that should be considered. One possibility is that an intervention changes the construct (e.g., increases prosocial behavior) and does not alter how the construct is measured (i.e., the measurement properties described above). This is the possibility that is often assumed but not tested. A second possibility is that an intervention leads to higher *measured* prosocial behavior but does not change the actual construct. In this scenario, the construct does not change, but only the measurement of it changes; for example, the intervention might cause some items to load more strongly and other items less strongly on the construct. In this case, researchers would erroneously conclude success of the intervention if they do not assess measurement equivalence. A third possibility is that an intervention increases prosocial behavior and also impacts the measurement of it so that the intervention effect is exaggerated. This exaggeration of the intervention effect hinders advancement of our understanding and provides overconfidence in applying the intervention in the future. A fourth possibility is that an intervention increases prosocial behavior but impacts the measurement process so that the intervention effect is hidden. Here, the benefit of the intervention goes undetected if measurement equivalence is not considered (this is also a possibility if the measures is not valid; i.e., does not measure the aspect of character impact by the intervention). A fifth possibility is that an intervention *reduces* prosocial behavior but heightens the reporting of prosocial behavior. Here, a harmful intervention appears to have no, or even a small beneficial, impact. It is currently impossible to know how common each of these possibilities are within the intervention research on character development and education. However, consistent evaluation and imposition of measurement equivalence in future studies would allow us to understand which possibilities are common, and to refine the measures used so as to ensure that the first possibility is assured rather than assumed.

Two Examples of Different Situations

Against this background information on evaluating the psychometric properties of measures and the relevance of using these techniques in the study of character development and education, I next present two very different situations from the measurement of different character strengths. The first situation is one in which an area of research is characterized by the frequent use of a limited number of measures; I use research on gratitude to illustrate this situation. The second situation is when there do not exist widely used measures,

and seemingly each study uses a different measure developed by the study authors. This latter situation is unfortunately more common in the study of character strengths, and I illustrate this situation with research on humility. These two examples come from ongoing work attempting to meta-analytically synthesize the psychometric properties of scales assessing various character strengths (Card, in progress). Although this work is ongoing and the full results of these meta-analyses are not presented here, the general states of these science of measuring gratitude and humility are illustrative as two examples of how research fields have progressed.

Situation 1: Limited Number of Widely Used Measures. The first situation is that in which a field is characterized by the presence of a limited number of widely used and well-established measures. In the ongoing meta-analyses of measures of character strengths, this situation was rare. However, the study of gratitude represents an example approaching this situation.

Gratitude can be defined as a sense of thankfulness or appreciation in response to receiving a gift, whether that gift is a tangible object given by someone else, experiences that one has had in life, or positive characteristics such as one's health (see e.g., Peterson & Seligman, 2004). The scientific study of gratitude has been influenced from its early stages by scientists concerned with measurement. As such, a few scales to measure gratitude were developed and evaluated early in this line of research. I next describe some strengths and limitations of two of these initial studies in order to give a sampling of this initial research.

McCullough, Emmons, and Tsang (2002) developed a six-item gratitude questionnaire known as the GQ-6 that has been widely used in subsequent gratitude research. Study 1 of McCullough et al. (2002) presents the first psychometric analysis: The authors administered a 39-item measure believed to assess gratitude to undergraduate students. Exploratory factor analyses indicated one predominant factor, and the authors trimmed the scale down to six items based on both conceptual and empirical (item-factor correlations) criteria. This study also assessed the correlation of self-reports of gratitude with reports of others (i.e., interinformant reliability) and with other self-report measures for which there were plausible expectations of correlations (i.e., construct validity). Study 2 administered the six-item measure to a larger sample of adults with a wider age range, and found similar evidence of construct validity. The other studies evaluated substantive questions involving gratitude using the six-item scale.

Watkins, Woodward, Stone, and Kolts (2003) developed the Gratitude Resentment and Appreciation Test, which is a multidimensional measure tapping various aspects of gratitude. Study 1 administered 55 items believed to assess gratitude to undergraduate students. Initial analyses led to 9 of these items being dropped to improve internal consistency. The authors offered initial expectations for a four-factor solution, but conducted exploratory factor analysis and empirically concluded a three factor solution. Study 2 of this article, also sampling undergraduates, used the factor structure indicated in Study 1 when evaluating test-retest reliability and construct validity. The remaining two studies included an experimental manipulation to impact gratitude, also using the three factor structure identified in Study 1.

Each of these seminal works is impressive both in translating theoretical ideas about gratitude into tractable measures and in the scope of studies in these early papers. Further, these papers collectively report many of the psychometric properties described above, including internal consistency reliability, interrater reliability, test-retest reliability, and numerous correlations informing construct validity. At the same time, these papers also contain limitations. Both studies removed items to improve reliability (though McCullough et al., 2002, also conceptually consider items retained). Both studies retained the empirically based decisions of items to retain (and factor structure, in Watkins et al., 2003) in subsequent

studies without reported replication. The studies relied on undergraduate samples in seven of the eight studies reported, and even the one noncollege sample were not ethnically diverse (91% White in Study 2 of McCullough et al., 2002). Validity evidence was drawn primarily from other self-report measures without considering the impact of shared-method variance. Finally, neither study considered or reported results of analyses of measurement invariance (across e.g., gender).

The purpose of these critiques is not to disparage these works. Both were novel and very rigorous efforts to develop new measures of gratitude. Instead, the purpose of these critiques is to point out that no studies are perfect, and it is critical for subsequent research to continue to evaluate psychometric properties of instruments rather that considering the psychometric properties established and not requiring further attention. Ongoing evaluation of psychometric properties should continue to inform potential modification of these instruments when necessary for a particular population, in a particular cultural context, or in different setting (thus extending the well-known specificity principle, as described in Bornstein, 2006, to specificity of measurement and measurement setting). These initial studies, and some other papers introducing additional measures of gratitude, have contributed to a field of research that has used these and a few other scales in a large number of other studies.

The ongoing review of measures of gratitude (Card, 2018) indicates that most studies in this area use one of four measures (including the two described above). Thus far, we have found at least 108 studies using and reporting internal consistency estimates from at least one of these four measures. Many studies use more than one measure, providing information regarding the overlap among measures. Although many of the studies consist of samples that by themselves are homogeneous, the collection of these 108 studies administers the measures in different countries and translated into various languages, and includes both basic and intervention studies.

The advantage of this situation in which many studies have used a small number of measure is that each measure has been used many times, so there are multiple prior studies from which to examine psychometric properties. Researchers considering a particular measure for a planned study have a sizable pool of prior studies from which to find psychometric information about how the measure has performed in similar populations, cultural contexts, and settings. However, it is notable that although most of the studies provide internal consistency reliability estimates, and some report correlations across different measures, the full range of psychometric information I described above is not typically reported. Therefore, the existing research on gratitude, despite the benefits of using a small number of measures in many studies, still suffers some of the limitations that the more general field of character development and education suffers that I mentioned above.

Before ending discussion of gratitude measures, I want to raise one more caution that could apply to any field using a small number of measures. The presence of a small number of widely used measures could potentially have negative ramifications for a field. The first possibility is that the operational definition of gratitude could evolve so that it is defined by what the prevailing measures are believed to measure. That is, the field might begin to move the bullseye (construct centroid) within Figure 1 to align with the midpoint of the indicators of commonly used scales, rather than continuing to refine the measures to better match a theoretically based operational definition. The existing measures of gratitude tend to have few items; although this is advantageous for minimizing time demands on participants, it is unlikely that these items capture the entire domain space of the measure of gratitude. In short, it is important for the field to continue to strive toward increasing validity rather than presuming that existing measures are the best possible. A second possibility is that researchers in a field could become dogmatic against the introduction of new measures. Despite the

range of countries and situations to which these measures have been applied thus far, researchers will likely continue to study gratitude in different contexts, in different cultures, and in different ages than have been studied thus far. As this extension of research occurs, researchers must have the flexibility to modify scales to suit emerging research needs.

Situation 2: Absence of Widely Used Measures. A very different situation is one in which there exists few or no widely used instruments of a construct of interest, so seemingly every researcher develops their own measure for every study conducted. In the ongoing meta-analysis of measures of character strengths, there were several areas of study that were in this situation. Here, I select the study of humility to illustrate these points.

Humility is a character strength that includes having an accurate sense of one's abilities and achievements, an ability to acknowledge mistakes, openness to advice and new ideas (Peterson & Seligman, 2002). Definitions of this construct vary, however, across publications and researchers, creating what was described above as the fuzzy boundary of operational definition.

In the ongoing meta-analyses, initial searches located a large number of studies of humility. However, a challenge was that many study authors used a measure of humility newly developed for that study. This multitude of different measures across studies created multiple challenges as a consumer of this research. First, one has to make decisions for each report whether the construct studied meets an operational definition of humility. Careful reading of this literature quickly showed that many reports labeled a construct as "humility" even though the items did not seem to capture the operational definition of the construct. Second, there also exists other studies using measures that appeared to meet the operational definition of humility, but those authors used a different label for that variable. It is impossible to know how many studies like that might exist, as there is no tractable way to search for these studies using electronic search engines (whereas if common terminology or measures were used, one could search for those terms). These first two challenges make it extremely difficult for consumers of research to find the relevant research on humility.

A third challenge caused by this multitude of measures is that, even if one can locate the relevant studies, it is extremely difficult to synthesize the psychometric properties in a way that is useful for planning future research. This is because there is simply insufficient use of a single measure across enough studies with varying populations, cultural contexts, and settings to allow for either adequate generalization of psychometric properties or an understanding of when measures perform better or worse. Instead, researchers face one of three poor choices:

The first choice is that researchers can find a prior study sampling a similar population, in a similar cultural context, and in a similar study, that shows adequate psychometric performance, and then use whatever measure was used in that research. This choice is problematic because the researcher is then constrained to use the measure of the prior study even if it does not fit their own operational definition of the construct. Research that pieces together measures of constructs from imprecise, sometimes overlapping operational definitions, often rooted in different theoretical perspectives, is simply a sloppy way to perform research and will produce inferential conclusions and estimates of effect sizes that are of little interest.

A second choice that researchers might make in the absence of synthesized psychometric evidence from prior studies is that they will choose the measure that best meets their operational definition, and look for evidence of psychometric adequacy in prior studies that differ in terms of the population studied, the cultural context, and/or the setting in which the measure was administered. Researchers making this choice wrongly assume that psychometric properties are characteristics of the measure, rather than point estimates of the

measure applied in a particular population and context. More technically, this choice ignores the fact that psychometric properties are point estimates subject to both population-level variability (due to sample, cultural context, setting, and innumerable other features) as well as sampling variability.

A third choice is simply to not use the prior literature to guide decisions in selecting measures of a construct. This choice might appear to be a nonsensical option, but it does appear to be a common choice in many studies of humility and other character strengths (at least, many reports do not provide a different rationale). Instead, a researcher making this choice develops a measure of the construct independently, creating items based on his or her own understanding of the construct, and perhaps even in a more systematic way such as preliminary qualitative studies. The researcher then administers the measure within the study, and performs analyses to further hone the measure. If these analyses to evaluate psychometric properties are well-performed, then the information is accurate and merely represents a less efficient way to accumulate information than if measures were used more consistently across studies. A worse case is that the psychometric analyses are not well-performed. These problematic analyses might include the practices described earlier, such as internal consistency analyses in which items are dropped if a higher Cronbach's α can be achieved, and sometimes exploratory factor analyses in which labels are applied post hoc if multidimensionality is found (i.e., the process of post hoc labeling factors itself is not problematic, but if the process becomes more an endeavor of creative explanation of findings than a dialogue of theory and data, the result is unlikely to generalize beyond the particular study). In the worst case, a researcher might be tempted to perform analyses using multiple ways of calculating the variable and select the method that provides the most favorable results (a highly problematic practice). I admit that the process described here is more problematic than is what likely actually occurs, but if it is plausible that the processes occur to some degree then a field of research is hampered. That is, if this is the typical practice in a field of research on character development and education, it is clear that comparison of results across studies is, at best, highly imprecise and, at worse, simply impossible.

PROPOSAL FOR A NEW APPROACH

In the previous section, two situations were described. The first situation was that a construct is measured frequently with a small number of instruments, but these scales have not been subject to extensive, ongoing evaluation of the full range of psychometric properties. The study of the character strength gratitude was offered as an example. The second situation was that a construct was not regularly assessed with any common measures, but instead each study used a different measure of the construct. I noted that several character strengths fit this situation, and offered humility as a specific example. Although I focused on only two character strengths based on a single, ongoing research synthesis, I challenge researchers across the broad field of character development and education to consider if the constructs that they study share the features of one of these two situations.

If either of the two situations that I have described, or some combination of both, characterizes the study of a construct relevant to character development and education, advancement in understanding is hampered. Although the second situation (i.e., little consistency in measures) is more problematic than the first situation (i.e., inadequate attention to the full range of psychometric properties of widely used measures), both are challenges to researchers in the field and to the broader accumulation of knowledge. In the remainder of this section, I offer suggestions that I believe offer solutions to many of these problems. Specifically, I offer suggestions to researchers planning studies, to researchers reporting study results, for research explicitly focused

on psychometrics of scales, and for efforts to synthesize existing research.

Suggestions for Study Planning. Researchers planning a study should be thoughtful consumers of prior research evidence of psychometric properties of scales. The first aspect of this recommendation is that researchers need to be consumers of prior research; to actively read and evaluate a large amount of the relevant research with the goal of selecting a best instrument for the planned research project. The second aspect of this recommendation is that researchers should be thoughtful in evaluating the prior research. I do not intend to imply that researchers are typically unthoughtful, or that only expert psychometricians are qualified to evaluate this literature. Instead, I wish to encourage researchers to keep in mind two points made earlier in this chapter: (a) that internal consistency reliability is no more important than other aspects of reliability, validity, or equivalence of measures; and (b) that all psychometric properties are point estimates based on how a measure was used, and one needs to consider similarities and difference in the sample, cultural context, and setting of the planned study relative to prior studies.

After evaluating previous studies of measures, I believe that researchers should be empowered to modify previously used scales. The dogmatic application of a particular scale implicitly assumes that psychometric properties are characteristics of the scale, whereas I believe that intelligent researchers who have thought carefully about the scale and the particular research setting are in the best position to decide how to measure a construct for their planned study. The selection of a measure and any modifications of it do need to be justified and well-considered, however. It is worth remembering that the most rigorously designed study that provides sophisticated answers about an uncertain construct has little value.

Suggestions for Reporting Findings. Once data has been collected for a study, researchers should fully report the psychometric properties of the measures of the study. I mentioned that Cronbach's α is a commonly reported index for internal consistency, and it is certainly expectable that researchers report this for each measure of a study. In addition to simply reporting Cronbach's α, however, researchers should consider if the stringent assumptions of this index are appropriate. Researchers should typically perform some sort of factor analysis, either exploratory if the number and structure of factors is entirely unknown, or else confirmatory in the more common case when the researcher has one or more ideas about the likely factor structure. The results of this factor analysis also provide item loadings which can be used to compute alternative, and potentially more appropriate, indexes of internal consistency reliability (e.g., McDonald's ω) that should also be reported.

Researchers should also consider psychometric properties beyond internal consistency reliability that can be reported. If available, correlations among multiple reporters (interinformant reliability) or repeated measures across a short timespan (test-retest reliability) should be reported.

In most studies, evidence for validity is available though not necessarily definitive. As described above, evidence for validity comes from the extent that the correlations of the *measure* with other variables corresponds to expectations of the correlations of the *construct* with the other variables. To the extent that expectations are met, this is evidence of validity; to the extent that correlations are not as expected, this is evidence against validity. A challenge is that most researchers do not focus on associations that have been investigated in enough previous studies to have solid expectations for correlations (given the pressure for novelty of research). Therefore, if correlations are not as expected, it is unclear if this is due to a lack of validity of the measure or because the initial expectations of the associations of the construct were inaccurate. One possibility to help overcome this challenge would be for researchers to add, in addition to the novel associations that motivate the study, additional

variables that have been widely studied and can be used for validity evidence.

Finally, many research studies could provide information about measurement equivalence, whether across gender, ethnicity, age, country, language of administration, format of administration (e.g., interview versus computer administration), treatment condition, setting, or any other aspect of variability contained in the study. Consistently evaluating, rather than simply assuming, whether measures function equivalently across participants in studies would provide a wealth of information for future research. This consistent evaluation would also respond to the need to explore the robustness of developmental findings (Duncan, Engel, Claessens, & Dowsett, 2014). The major challenge to doing so, however, is that the data analyses needed to properly evaluate measurement equivalence (i.e., confirmatory factor analysis or item response theory) are more advanced than some researchers would otherwise use for their studies, and might be inaccessible or just too time-consuming for some researchers. However, for projects that are well-supported and/or have a dedicated data analyst, the practice of evaluating and reporting measurement equivalence would be valuable enough to justify the small amount of additional resources.

Suggestions for Studies Focused on Psychometrics. Some of the recommendations offered in planning and reporting studies require deliberate attention to psychometrics. Extending this recommendation further, it would be valuable for some studies to explicitly focus on evaluating the psychometric properties of measures. Such studies might pit multiple measures of a construct head to head, and compare performance across a wide range of sample characteristics, settings, and/or cultural contexts. These studies would likely include additional variables to assess validity, and have a plan for evaluating equivalence of measures.

Studies focused exclusively on evaluating psychometric properties of measures are rare. This rarity might be in large part because it is difficult to find publication homes for such studies. It is this author's perception that many journals are less likely to publish studies focused on psychometric properties than other types, based on the perception that psychometric studies are useful for next steps in research, but by themselves as not advancing understanding. I believe that this perception is both incorrect and shortsighted. It is incorrect because knowing how a construct is best measured—such as, how various items believed to represent part of the construct domain overlap with each other and other constructs, and how this system of indicators perform differently across populations and contexts—does advance our conceptual or theoretical understanding of the construct itself. Further, this perception seems shortsighted to this author because there is little advancement of knowledge of a construct from studies that lack evidence that they are actually measuring a construct well. Changing perceptions toward a view that studies of psychometric properties represent the foundation on which further research can be performed are needed. Researchers conducting, justifying, and demonstrating the importance of these efforts would help change this perception.

Suggestions for Research Accumulation. In addition to conducting studies specifically designed to assess psychometric properties of measures, another valuable effort toward acquiring information on these properties would be to systematically review the existing research. Earlier I suggested that researchers planning a study consult previous literature to obtain estimates of psychometric properties from similar studies. However, I noted that it can be challenging to find previous studies that closely match the planned study in all relevant ways (e.g., population studied, cultural context, setting). Further, conducting an extensive search is time consuming, and although this is not an excuse to not review the literature, the reality is that many researchers will not necessarily have the time and resources to conduct an extensive and careful review. Therefore, an effort to summarize the available research

could be very beneficial to researchers planning studies.

In some cases, a synthesis of prior studies might be advantageous even compared to a study designed to evaluate psychometric properties. All studies are limited in some ways, such as limits to sample sizes, limits to populations that can be accessed, limits to the number of measures or administration methods possible, limits to the settings in which participants are drawn, and limited to the country or cultural context studied. In some fields, it may be more effective, and certainly more economical, to synthesize the psychometric properties from existing studies rather than (or in addition to) a more limited study of the psychometric properties.

If a research synthesis of the psychometric properties of measures is conducted, then meta-analysis offers a useful methodology for conducting this synthesis. Broadly construed, the term meta-analysis refers to the formulation of research questions about effects across studies, thorough literature searches to find relevant studies, systematic coding of study characteristics and effect sizes, statistical analyses of those effect sizes and coded study characteristics, and techniques of presenting these results (Cooper, 2009). Full details of the techniques of meta-analysis are beyond the scope of this article (for details, see Card, 2012; Cooper, Hedges, & Valentine, 2009; Lipsey & Wilson, 2001; Rosenthal, 1991). Here, however, I note that meta-analysis can serve as an effective tool for aggregating research findings to obtain more precise estimates of psychometric properties, and can also be used to identify the conditions under which measures are expected to have differing psychometric properties. For a more complete description of the use of meta-analysis for research accumulation, see Card (in press).

Although meta-analysis offers a useful methodological tool for this type of psychometric synthesis of measures of character, there are multiple challenges of this approach. First, this approaches requires an adequate number of studies that use a common, or at least similar, measure of the construct. I described that the research on gratitude fits this requirement, whereas the research on several other character strengths, including humility, does not. Second, this approach requires that psychometric properties are reported for each study (or can be obtained from study authors). A surprisingly large number of studies of the character strengths relevant for the ongoing meta-analyses mentioned earlier did not report internal consistency, and very few reported other psychometric properties even when it appeared that the data could have been available. Third, the studies that exist should be of sufficient quality and variability to make the combination of psychometric properties yield a meaningful result. If all studies are of questionable quality (e.g., unclear sampling frame) or the measures included don't match well the operational definitions of constructs of interest, then sophisticated combination of results still leads to information of minimal value. Finally, in order to provide meaningful information about the conditions in which a measure works better or worse, there must be variability across studies in the conditions in which the measure was used. For example, if a measure has been applied only with adult samples, it is impossible for meta-analysis of those studies to inform whether the measure performs well with adolescent samples. Because the study of character strengths is relatively new, many measures have been used by a limited number of researchers in a limited number of settings. It is worth considering the variability of studies of any measure of character development and education before planning a meta-analysis of those studies.

CONCLUSIONS

In this article, I have offered several suggestions for improving methodology in the measurement of character development and education. First, I have encouraged attention to

all three psychometric qualities of measures, considering validity and measurement equivalence as well as reliability. Second, I have emphasized that evaluation of psychometric properties, and modifying measures based on these results, is an ongoing process. Best practices in measuring character, character development, or character education should be an ongoing process of an active science. Third, I argued that all studies should attend to psychometric properties (i.e., psychometric considerations are not the exclusive domain of methodologists). Researchers planning studies should select measures of constructs based on previous psychometric evidence, matched as closely as possible to the specific sample, context, and goals (e.g., measurement of stable constructs versus sensitivity to change across time) of the planned study. Researchers reporting study results should also report the psychometric properties—that is, evidence of reliability, validity, and equivalence described in this paper—found in the study. Even if these psychometric properties are not the main motivation for conducting a study, full reporting is critical to building a science in which a repository of psychometric information is available. Finally, I have suggested dedicated efforts of research synthesis to create an ongoing source of information to summarize what is known about measuring character development and education.

I hope that the considerations I have offered provide a useful foundation for advancing measurement of character development and education. I believe that, while not the only important aspect, an informed and consistent attention to measurement is critical for the scientific study of character development and education.

Acknowledgment: An earlier version of this article was prepared in conjunction with the National Academy of Science workshop on Character Development, July 26–27, 2016. The author gratefully acknowledges support for this work by a grant from the John Templeton Foundation (IF#47910).

REFERENCES

Bornstein, M. H. (2006). Parenting science and practice. In K. A. Renninger & I. Siegel (Volume Eds.) and W. Damon & R. M. Lerner (Eds.), *Handbook of child psychology: Vol. 4. Child psychology in practice* (pp. 893–949). Hoboken, NJ: Wiley.

Campbell, D. T., & Fiske, D. W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin, 56*, 81–105. http://dx.doi.org/10.1037/h0046016

Card, N. A. (2012). *Applied meta-analysis for social science research*. New York, NY: Guilford.

Card, N. A. (2013). Psychometric considerations in cyberbullying research. In S. Bauman, D. Cross, & J. Walker (Eds.), *Principles of cyberbullying research: Definitions, measures, and methodology* (pp. 188–201). New York, NY: Routledge.

Card, N. A. (Ed.) (2017). Developmental methodology. Monographs of the Society for Research in Child Development, 82(2).

Card, N. A. (2018). *What is known about existing measures: Meta-analyses of psychometric properties of measures of twelve character strengths*. Manuscript in preparation.

Card, N. A., & Little, T. D. (2006). Analytic considerations in cross-cultural research on peer relations. In X. Chen, D. C. French, & B. Schneider (Eds.), *Peer relations in cultural context* (pp. 75–95). New York, NY: Cambridge University Press.

Cheung, G. W., & Rensvold, R. B. (2002). Evaluating goodness-of-fit indexes for testing measurement invariance. *Structural Equation Modeling, 9*, 233–255. doi:10.1207/S15328007SEM0902_5

Cooper, H. M. (2009). *Research synthesis and meta-analysis: A step-by-step approach*. Thousand Oaks, CA: SAGE.

Cooper, H., Hedges, L. V., & Valentine, J. C. (Eds.). *The handbook of research synthesis and meta-analysis* (2nd ed.). New York, NY: Russell Sage Foundation.

DeVellis, R. F. (2003). *Scale development: Theory and application*. Thousand Oaks, CA: SAGE.

Duncan, G. J., Engel, M., Claessens, A., & Dowsett, C. J. (2014). Replication and robustness in developmental research. *Developmental Psychology, 50*, 2417–2425. doi:10.1037/a0037996

Lipsey, M. W., & Wilson, D. B. (2001). *Practical meta-analysis*. Thousand Oaks, CA: SAGE.

Little, T. D. (1997). Mean and covariance structures (MACS) analyses of cross-cultural data: Practical and theoretical issues. *Multivariate Behavioral Research, 32*, 53–76. doi:10.1207/s15327906mbr3201_3

Little, T. D., Lindenberger, U., & Nesselroade, J. R. (1999). On selecting indicators for multivariate measurement and modeling with latent variables: When "good" indicators are bad and "bad" indicators are good. *Psychological Methods, 4*, 192–211. doi:10.1037/1082-989X.4.2.192

McCullough, M. E., Emmons, R. A., Tsang, J.-A. (2002). The grateful disposition: A conceptual and empirical topography. *Journal of Personality and Social Psychology, 82*, 112–127.

McDonald, R. P. (1999). *Test theory: A unified treatment*. Mahwah, NJ: Erlbaum.

Millsap, R. E., & Cham, H. (2012). Investigating factorial invariance in longitudinal data. In B. Laursen, T. D. Little, & N. A. Card (Eds.), *Handbook of developmental research methods* (pp. 109–127). New York, NY: Guilford.

Nunnally, J. C. (1978). *Psychometric theory*. New York, NY: McGraw Hill.

Peterson, C., & Seligman, M. E. P. (2004). *Character strengths and virtues: A handbook and classification*. New York, NY: Oxford University Press.

Rhemtula, M., Brosseau-Liard, & Savalei, V. (2012). When can categorical variables be treated as continuous? A comparison of robust continuous and categorical SEM estimation methods under suboptimal conditions. *Psychological Methods, 17,* 354–373.

Rosenthal, R. (1991). *Meta-analytic procedures for social research* (Rev. ed.). Newbury Park, CA: SAGE.

Watkins, P. C., Woodward, K., Stone, T., & Kolts, R. L. (2004). Gratitude and happiness: Development of a measure of gratitude, and relationships with subjective well-being. *Social Behavior and Personality, 31*, 431–452. doi:10.1037/t00741-000