In the manuscript entitled, *A Test of the Adequacy of Three Common Core Youth Measures with Latino Immigrant Youth*, Liu and colleagues (2017) examine the important issue of measurement equivalence – the degree to which a measure assesses the same construct in the same way across population groups (Knight, Roosa, & Umaña-Taylor, 2009). The authors, in particular, examine measures from an emerging area of research, Positive Youth Development (PYD), to determine the psychometric properties of character and caring for use with Latino immigrant youth. This is a critical methodological topic that has not received due attention by content scholars working with ethnic minority, international, or economically disadvantaged populations.

In the spirit of the TCRM workshops, my goal in the following comments is to challenge us to consider the role of methodology in our understanding of family science. Given the aim of the current paper to consider the adequacy of the measures of character and caring for use with Latino immigrant youth, I would like to devote some of my comments to the significance of measurement issues in family research. Measurement is a critical component of our science as many constructs must be examined using approximations and assignment of numerical values to represent participant’s status on those constructs. Measurement is fundamental in determining the internal and external validity of our findings. As many of our measures have been developed with the use of predominantly White, European American middle class samples, measurement validity becomes ever more important as we move to doing research with other populations with
different backgrounds and experiences (e.g., ethnic minority, ethnically disadvantaged) (Knight et al., 2009). Often researchers, do not give thought to whether measures are reliable and valid or even appropriate to use with, for example, ethnic minority samples. One might ask, do scores on measures developed with members of a population who are individualistic convey the same meaning with different populations, for example, that are collectivistic in nature. How does one even go about addressing this question without it just being an afterthought?

In Liu et al.’s paper, the authors present a review of some of the methodological issues when considering the use of measures with different populations; I commend the authors for taking on this important task in their paper. In many ways, the authors took important steps in bringing to light the important and complex issue of measurement equivalence across varying populations. However, I found myself questioning how well the analysis chosen (i.e., longitudinal temporal stability) addressed this important question put forward by the authors. I encourage the authors to consider the match between their research question and the methods used to answer their question. In particular, Liu and colleagues argue that there is a need to test for the ‘construct adequacy’ of the PYD constructs of character and caring as being salient for Latino youth. If this is the question that Liu and colleagues are trying to address, I wondered if the tests for longitudinal temporal stability are premature. Discussions of measurement equivalence across different populations (e.g., Knight et al., 2009; Millsap, 2011) suggest a particular sequence, for example, (a) articulating a theoretical understanding of critical aspects of the construct (or expected variations in the construct) for the particular population of interest (which Liu et al., have started to lay out), (b) item identification (or a review of the items developed with another population), (c) factor analysis – does the measure have a similar factor structure in relation to other published factor structure studies with other populations, (e) identifying how the construct
of interest might relate to other theoretical constructs of interest (to start examining validity relations – covariation with scores produced by those theoretically related measures; convergent and discriminant validity), and so on. Based on this delineation of steps in relation to measurement equivalence, I would argue that prior to taking on analyses of longitudinal temporal stability, there needs to be evidence that the measure understudy assesses the construct in the same way with the new population group (i.e., construct validity). If not, then how do we know that the measure is not consistently measuring the wrong thing over time? In addressing questions of measurement equivalence, I would encourage authors to include a review of all evidence related to the measure with the population of interest in relation to the measurement steps as reviewed above (e.g., theory, item review, factor structure, reliability and measurement error, convergent and discriminant validity).

In Liu and colleagues’ paper, they have articulated important differences related to the population that the PYD measures were developed with as compared to their population of interest, Latino youth. I ask them to take this a bit further to strengthen the theoretical foundation in relation to the PYD constructs under study for the Latino population. First, they have presented a discussion of cultural variation, such as individualism as compared to collectivism, and how these constructs might be important when considering positive youth development. I would encourage the authors to take this a step further when thinking about the question of measurement equivalence and consider/articulate, might they expect variation in response patterns related to these important aspect of Latino culture in relation to the population within which the measure was developed and what might that variation look like? Second, the authors have pointed out the tremendous variability within the Latino population that might result in differences between the Latino population and other populations and within the Latino
population. Again, taking this a bit further, might the authors expect variation in response patterns in relation to the measures under study; and what might this look like? Do they expect all of the items included in each measure to work as effectively as they do other measurement studies of the PYD constructs? I would challenge the authors to consider what hypotheses they might have about potential theoretical differences in the factor structure of caring and compassion for the Latino population. If they expect within group variability (e.g., by acculturation, language use), might there be a need for the examination of invariance across levels of these moderator variables if the sample includes this variation? I encourage the authors to detail this rationale within the introduction (e.g., theoretical support for measurement equivalence, items and factor structure, validity tests, etc.); this should match what is presented in the methods and results sections.

An important issue related to research methodology more broadly, is the presentation of our methodological and analytic choices in the context of manuscripts. With issues of replication and transparency in science, I encourage authors to be as detailed as possible with their methodological details. The current manuscript, in particular, could serve as an empirical example for readers interested in measurement equivalence across populations; thus, I would encourage the authors to provide additional details in this section to help readers follow the rationale of their methodological and analytic choices. Here I highlighted a few places that would benefit from additional details. First, in the authors’ presentation of the Analytic Strategy section, I wondered how their analysis of the normal distribution of the items compared with prior work on these measures? Have other researchers found similar issues with these items or is this unique to the current sample? Based on the authors’ presentation of the response choices, they seem to conform to typical Likert measures of psychosocial constructs. Did they consider
using full information maximum likelihood – robust in their Mplus analysis; conversely, I ask the authors to further articulate why they choose WSLMV? Second, I was curious as to why the authors chose multiple group confirmatory factor analysis (i.e., data stacked by time); I would ask the authors to detail how they handled the nested nature of the data (i.e., time nested within individuals)? Third, I ask that authors to provide a more detailed rationale for why they only test configural and scalar invariance models? For example, in work by Little (2013) on longitudinal factor analyses, he recommends including distinct latent factors for each time point (i.e., not stacking the data), as well as four stages of tests for longitudinal factorial invariance (i.e., configural to strict factorial invariance). Fourth, the authors state that they used a chi-square difference test, but they do not report the chi-square values in the manuscript (i.e., Table 1 – how was the DIFF Test calculated? Was the chi-square value used for this?) and do not list the chi-square statistics as one of their model fit indices. Lastly, it would be helpful if the authors could present the parameters from the measurement model either in a table or a figure (for recommendation on CFA reporting practices see Jackson, Gillaspy, & Purc-Stephenson, 2009).

I look forward to discussing broader methodological issues as framed by this paper.
References


