

Commentary: The Critical Role of Measurement (and Space Elevators) in the Study of Child Development

Patrick J. Curran, PhD

University of North Carolina at Chapel Hill

All correspondence concerning this article should be addressed to Patrick J. Curran, PhD, Department of Psychology, University of North Carolina, Chapel Hill, NC, 27599-3270, USA. E-mail: curran@unc.edu

Received December 19, 2013; revisions received December 19, 2013; accepted December 23, 2013

Key words assessment; statistical applications; structural equation modeling.

A few years back I became friends with a research scientist who was a member of a molecular physics lab on campus. We mostly spent our time arguing about baseball (the pitcher should *always* bat), but occasionally we would talk about work. He was researching the development of a space elevator that consisted of a satellite placed in geosynchronous orbit while tethered to the ground, allowing an elevator to deliver payloads to space. He would talk about things like nanotubes, microfabrication, noncovalent intermolecular forces, and molecular assemblers, and then would ask about my own current projects. After hearing about his work, I would sheepishly tell him about how I was trying to develop a measurement model for childhood depression. I described how we had gathered binary data indicating the presence or absence of symptoms obtained from a sample of children followed over time and that we were trying to map the binary items onto a continuously distributed score to model trajectories of depression. I felt so sheepish because our empirical data were nothing more than a rectangle of zeros and ones, the underlying mathematics did not go much past high school calculus, and the requisite computer programming constituted a few dozen lines of code at best. Yet at the end of my description my friend laughed, shook his head, and said that he was glad that he had the easier problem on which to work; he had no desire to tackle something as challenging as trying to measure depression in children.

It may seem curious that the scientist who owns the molecular assembler is happy that he does not have to work with the complexities of a data matrix of zeros and ones, but this is precisely the situation in which we find ourselves when studying the development of children over

time. This is exemplified in the remarkable set of papers that make up this special issue on quantitative methodology. This corpus of work offers a clearly articulated and broadly accessible tour of the striking advances that have been achieved in statistics and methodology over the past few years. Topics addressed include receiver operating characteristic analysis, multilevel survival analysis, missing data, mediation and moderation, single-case research designs, latent variable mixture models, n-of-1 randomized controlled trials, and advanced structural equation models of change. Further, these methods are used to study important developmental phenomena such as internalizing symptomatology, parental influences on child behavior, adolescent achievement goals, social norms and intentions, and coping with pain, among many others. Despite the myriad of statistical methods used and the corresponding array of developmental questions under study, there remains one common thread that runs through the entire set of papers: *measurement*.

Nearly every developmental researcher has had to grapple with the fundamental challenge of measurement, namely, that which we believe to exist is test *independent*, yet the empirical data we obtain are test *dependent*. We work in a field of science in which rarely if ever are we able to directly observe what we most want to measure. For example, developmental theory describes the course, causes, and consequences of individually varying trajectories of depression (e.g., Graber & Sontag, 2009). However, to empirically estimate these trajectories we must first obtain a numerical measure of depression. To do so, we might ask a parent to report on whether their child “cried easily,” “was lonely,” or “was sad” during the prior 30

days. We must then somehow optimally combine the set of “yes” and “no” responses to obtain a valid and reliable measure of what theory describes as *depression*. Our measure hopefully captures a large part of the child’s depression, but it also likely contains elements of the parent’s own depression, the parent’s perceptions about what is typical child behavior, and countless other contributions that all comingle with true depression. Yet this numerical measure is all that we have available with which to test our research hypotheses.

Perhaps one of the most elegant and refreshingly concise definitions of measurement was offered more than half a century ago by Stevens (1946) who said that measurement is “. . . the assignment of numerals to objects or events according to rules” (p. 677). A simple example is that if a parent perceives that his/her child cried easily during the prior 30 days, the item is assigned a value of 1, else it is assigned a value of 0; we have followed a rule to assign a number to an object. Once these numerical values are obtained, we can then model these in a variety of powerful and flexible ways. For example, drawing on papers making up this special issue, Karazsia, Berlin, Armstrong, Janicke, and Darling (2014) expresses child body dissatisfaction as a function of maternal encouragement to diet; Berlin, Parra, and Williams (2014) expresses body mass index as a function of age; and Barker, Rancourt, and Jelalian (2014) express fear of negative evaluation as a function of being overweight. Every paper in this special issue builds some form of a model to express its outcome measures of interest.

However, a fundamental condition underlying all of these complex expressions is that the numbers we assign following our adopted rules of measurement validly capture our underlying theoretical construct for *all* children belonging to *all* subgroups at *all* ages. The classical term for this condition is *measurement invariance* (e.g., Meredith, 1964). Measurement invariance implies that the model that relates our set of items to the theoretical construct of interest does not meaningfully vary across individuals or over time, that is, each item equally defines the underlying construct across all subgroups of children at all ages. This is a daunting condition to achieve under even the best of circumstances, and it is made that much harder when children have the audacity to develop and change over time. Returning to our prior example, measurement invariance prompts us to ask ourselves if the item “cried easily” means the same thing for boys and girls, for children who are 4 or 8 or 12 years old, for children with or without a mother who are depressed, or whether the item was assessed at home or at school. Our traditional measurement models almost always assume that this item is equally indicative of depression regardless of age or subgroup

membership. However, the extent to which this invariance condition is not met directly undermines our ability to make valid inferences from our data (Shadish, Cook, & Campbell, 2002).

Although as a field we are ostensibly aware of the assumptions of invariance, I am becoming increasingly convinced that we do not provide adequate heed to this issue. There are of course exceptions to this rule in the study of pediatric psychology. For example, Ferro and Boyle (2013) rigorously tested longitudinal invariance in global self-concept for adolescents with and without a chronic illness, and Jekauc, Voelkle, Wagner, Mewes, and Woll (2013) examined measurement invariance of the physical activity enjoyment scale in children and adolescents across age and gender. However, my concerns primarily stem directly from my own work.

As I continue to collaborate with colleagues to try to better understand the etiological mechanisms underlying the development of substance use in children and adolescents, I have become increasingly cognizant of how difficult it is to develop measures that are truly invariant across subgroup and over age. The main challenge is that we *expect* children to change over time; this is of course the whole point of studying development. Thus, although we believe the theoretical construct of depression exists in some constant form across development, at the same time we also expect that how a child *expresses* depression will change across development, and these developmental changes are in addition to differences that we equally expect across subgroups such as gender, race, or diagnostic status.

But here lies what may be the most vexing problem we face: If measurement does structurally differ over group or across time, and we do not account for such changes in our measurement models, then it is likely that we will draw biased conclusions about what we believe to be true developmental change. This is because our fitted models are simultaneously trying to capture true developmental change and failed measurement invariance, the results of which do not accurately represent either process. The inextricable mixing of change in the true underlying construct with change in the measurement of the underlying construct represents a clear and present danger to the internal validity of our conclusions about developmental etiology and causal processes (Shadish et al., 2002). I believe this issue is currently one of the greatest challenges we face in developmental science.

My hope is that as a field we continue to enthusiastically pursue the exciting analytic strategies that have been so clearly demonstrated in this special issue, but at the same time make serious efforts to pay sincere attention to issues of

measurement invariance and construct validity. The most complex and rigorous statistical models are constrained by the extent to which we are truly measuring what we believe we are measuring. There are classic methods available for evaluating measurement invariance in both factor analysis (e.g., Cheung & Rensvold, 1999; Meredith, 1993; Widaman, Ferrer, & Conger, 2010) and item response theory models (e.g. Thissen & Wainer, 2001), as well as methods for incorporating potential measurement differences into score estimation (e.g., Curran et al., 2008; Flora, Curran, Hussong, & Edwards, 2008). There is also promising recent work on moderated nonlinear factor models (e.g., Bauer & Hussong, 2009) and multiple reporter models (Bauer et al., 2014) that extend these classical methods even further. Our field can only stand to benefit from the continued incorporation of these classic and contemporary approaches to measurement as we endeavor toward the future.

In conclusion, I believe this special issue highlights two particularly salient points. First, it is clear that novel methodologies are being applied to important developmental questions in new and exciting ways that allow us to empirically test theory with a flexibility and rigor that was not possible even a few years ago. Second, this collection of papers even further exemplifies my friend's opinion that the study of child development can indeed be more complex than building a space elevator.

Funding

This work was partially funded by Award Number R01DA015398 (Patrick Curran and Andrea Hussong, co-PIs).

Conflicts of interest: None declared.

References

- Barker, D. H., Rancourt, D., & Jelalian, E. (2014). Flexible models of change: Using structural equations to match statistical and theoretical models of multiple change processes. *Journal of Pediatric Psychology, 39*, 233–245. doi:10.1093/jpepsy/jst082
- Bauer, D. J., Howard, A. L., Baldasaro, R., Curran, P. J., Hussong, A. M., Chassin, L., & Zucker, R. (2014). A tri-factor model for integrating ratings across multiple informants. *Psychological Methods, 14*, 101–125. doi:10.1037/a0015583
- Berlin, K. S., Parra, G. R., & Williams, N. A. (2014). An Introduction to latent variable mixture modeling (part 2): Longitudinal latent class growth and growth mixture models. *Journal of Pediatric Psychology, 39*, 188–203. doi:10.1093/jpepsy/jst085
- Cheung, G. W., & Rensvold, R. B. (1999). Testing factorial invariance across groups: A reconceptualization and proposed new method. *Journal of Management, 25*, 1–27.
- Curran, P. J., Hussong, A. M., Cai, L., Huang, W., Chassin, L., Sher, K. J., & Zucker, R. A. (2008). Pooling data from multiple longitudinal studies: The role of item response theory in integrative data analysis. *Developmental Psychology, 44*, 365–380. doi:10.1037/0012-1649.44.2.365
- Ferro, M.A., & Boyle, M.H. (2013). Longitudinal invariance of measurement and structure of global self-concept: A population-based study examining trajectories among adolescents with and without chronic illness. *Journal of Pediatric Psychology, 38*, 425–437.
- Flora, D. B., Curran, P. J., Hussong, A. M., & Edwards, M. C. (2008). Incorporating measurement non-equivalence in a cross-study latent growth curve analysis. *Structural Equation Modeling, 15*, 676–704.
- Graber, J. A., & Sontag, L. M. (2009). Internalizing problems during adolescence. In R. Lerner, & L. Steinberg (Eds.), *Handbook of adolescent psychology, Vol 1: Individual bases of adolescent development* (3rd ed., pp. 642–682). Hoboken, NJ: John Wiley & Sons Inc.
- Jekauc, D., Voelkle, M., Wagner, M.O., Mewes, N., & Woll, A. (2013). Reliability, validity, and measurement invariance of the German version of the Physical Activity Enjoyment Scale. *Journal of Pediatric Psychology, 38*, 104–115.
- Karazsia, B. T., Berlin, K. S., Armstrong, B., Janicke, D., & Darling, K.W. (2014). Integrating mediation and moderation to advance theory development and testing. *Journal of Pediatric Psychology, 39*, 163–173. doi:10.1093/jpepsy/jst080
- Meredith, W. (1964). Notes on factorial invariance. *Psychometrika, 29*, 177–185.
- Meredith, W. (1993). Measurement invariance, factor analysis and factorial invariance. *Psychometrika, 58*, 525–543.

- Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). *Experimental and quasi-experimental designs for generalized causal inference*. New York, NY: Houghton Mifflin Company.
- Stevens, S.S. (1946). On the theory of scales of measurement. *Science*, 103, 677–680.
- Thissen, D., & Wainer, H. (Eds.), (2001). *Test scoring*. Hillsdale, NJ: Lawrence Erlbaum Associates.
- Widaman, K. F., Ferrer, E., & Conger, R. D. (2010). Factorial invariance within longitudinal structural equation models: Measuring the same construct across time. *Child Development Perspectives*, 4, 10–18.