Measurement in the Context of Intervention

Mark White
University of Oslo

Author note

Correspondence concerning this abstract should be addressed to Mark White, Frognerveien 57, Oslo 0266. Norway. E-mail: mrkwht@umich.com

Context:

Imagine giving young students a survey on their ability to perservere on long term projects. Most students have had no real experience working on a project that lasts longer than a day so they respond to the survey based on what they believe that they might do when working on a long-term project. The school then conducts an intervention to promote students ability to perservere, which involves giving students multiple multi-week projects to work on. After the intervention, students complete the same survey, but now, by virtue of the intervention, they are able to self-report on how well they actually perservered on the projects. What was a survey that captured self-perceptions regarding one's response to novel situations now has become a survey capturing lived experiences. The very nature of the construct being measured changed due to the existence of the intervention, rendering meaningless changes in the scale created by the survey. This is an example of instrumentation error (Shadish, Cook, & Campbell, 2002). As schools start to implement large-scale measurement of social-emotional competencies (SEC) and actively seek to promote these competencies, it is important to take stock of this threat of instrumentation and develop approaches to detect and understand when and if it is occuring. The field, to this point, has focused on error due to reference bias and other similar errors (e.g. Heine, Lehman, Peng, & Greenholtz, 2002), but has no standard methods to try and address the instrumentation errors just raised. This paper proposes examining the nomological network (Cronbach & Meehl, 1955) around a measured construct both pre- and post- intervention as a way of understanding if the intervention changed the nature of the construct being measured.

Purpose:

The goal of this paper is to discuss threats that can arise when conducting measurement in the context of interventions. The fact that interventions can change the nature of the construct being measured is under-appreciated. This paper seeks to demonstrate this problem empirically and propose an approach to test for this problem using item-response theory.

Data:

Data comes from over 600,000 students who completed the Leader in Me Measureable Results Assessment (MRA). This assessment is designed to measure students' development of the 7 Habits, a set of leadership skills that are the main target of intervention by the Leader in Me program. Surveys were completed by students both in the first year of their schools participation in the program (before any intervention occurred) and in each subsequent year of intervention. While answering questions about the 7 Habits, students also responded to a number of other scales, including School Connectedness (Resnick, et al, 1997), Growth Mindset (Dweck, Chiu, & Hong 1995), Social Responsibility (Bandura, 1989), and others. These secondary scales were selected becaue they are more standard measures of similar constructs as the 7 Habits, btu are not targets of the Leader in Me intervention. All data is collected anonymously. Three forms of the survey were administered with students randomly assigned to a form and 7 Habit questions randomly distributed across forms. Secondary scales appeared only on one form.

Research Design:

The analysis takes the form of a multi-group analysis of measurement invariance using item response theory (IRT). The base group is pre-intervention students who took the survey in beform being exposed to the Leader in Me intervention. The focal group is students in the same schools who have been exposed to the Leader in Me intervention. The analyses focus on the invariance the measurement properties of the scales change across groups.

Using multi-group graded-response IRT models, I first used unidimensional IRT models to identify items that were invariant across groups. Restricting item characteristics for these items to be equal across groups, I then estimated multi-dimensional models using the 7 Habits items to form one dimension and secondary scales as other dimensions. The main test for the multi-dimensional models was whether restricting the covariance between latent constructs to be equal across groups led to better fitting models. If restricting the covariance to be equal across group led to better fitting models, it implies the nomological network of the 7 Habits construct is equal across groups. This would support the argument that the 7 Habits scale is measuring the same construct both for students not exposed to Leader in Me and those who are exposed to Leader in Me.

Results:

Analyses are still ongoing as this is being submitted as a work-in-progress. However, initial results suggest only a most items are invariant across groups. That is, only less than 10 of the 63 items from the 7 Habits scale showed differential item functioning. All items from secondary scales were free of differential item functioning problems. Results from multi-dimensional models are not yet available. However, the existing results show that any problems that might exist with the 7 Habits scale are not detectable by common approaches testing for non-invariance of items.

References

Bandura A. (1989). The Multidimensional Self-Efficacy Scales. Unpublished test, Stanford University, Stanford, CA.

Cronbach, L. J., & Meehl, P. E. (1955). Construct Validity in Psychological Tests. *Psychological Bulletin*, *52*(4), 281–302. https://doi.org/10.1037/h0040957

Dweck, C. S., Chiu, C., & Hong, Y. (1995). Implicit Theories and Their Role in Judgments and Reactions: A Word From Two Perspectives. *Psychological Inquiry*, *6*(4), 267–285. https://doi.org/10.1207/s15327965pli0604_1

Heine, S. J., Lehman, D. R., Peng, K., & Greenholtz, J. (2002). What's wrong with cross-cultural comparisons of subjective Likert scales?: The reference-group effect. *Journal of Personality and Social Psychology*, *82*(6), 903–918. https://doi.org/10.1037/0022-3514.82.6.903

Resnick, M. D., Bearman, P. S., Blum, R. Wm., Bauman, K. E., Harris, K. M., Jones, J., … Udry, J. R. (1997). Protecting Adolescents From Harm: Findings From the National Longitudinal Study on Adolescent Health. *JAMA*, *278*(10), 823–832. https://doi.org/10.1001/jama.1997.03550100049038

Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). *Experimental and quasi-experimental designs for generalized causal inference*. Houghton Mifflin.