47 NOT ESTABLISHING THE CROSS-CULTURAL VALIDITY OF MEASURES OF KEY CONSTRUCTS IN A HIGH-STAKES FIELD EXPERIMENT

J. Lawrence Aber
New York University

One of the most important tasks in conducting psychological research is to establish that the measures one uses in research reliably and validly index the constructs one intends to measure. Examples of constructs and measures in my subfield of developmental psychology abound. For example, the security of young children's attachment to their primary caregivers (the construct) is best measured by the pattern of young children's behaviors before, during, and after a brief separation from their parents in the "strange situation" paradigm (the measure; Ainsworth, 1970). Similarly, preschool children's ability to delay gratification (the construct) is measured by minutes of time they wait for a preferred reward in the "Marshmallow Test" (the measure) (Mischel, Shoda, & Rodriquez, 1988). But how do we know if measures truly index the constructs we wish to investigate? The two key ways are by establishing the measures' reliability and validity. The reliability of a measure is an index of the consistency of responses to the measure. The validity of a measure is an index of the extent to which it is measuring what it is supposed to measure.

The biggest research mistake I've made recently was to rely on evidence demonstrating the reliability of several measures of school-aged children's "self-regulation" as adequate to support the validity of the measures. I made this mistake by using the measures in the new cultures of Lebanon and Niger where the measures have never been used before. Let me explain the research context, the mistake, and its scientific and real-world implications.

For the last several decades, I have been conducting field experiments of school-based interventions to improve both the social-emotional development and academic learning of children. I began this work with studies in low-income and/or conflict-affected communities in the United States (Aber, Brown, Jones, & Roderick, 2010; Aber, Brown, Jones, Berg, & Torrente, 2011; Jones, Brown, & Aber, 2011), but starting in 2010, I began similar work in low-income and/or conflict-affected countries

(e.g., Democratic Republic of Congo, Niger, and Lebanon). My colleagues and I view such research as opportunities both to learn what interventions work best to promote children's learning and development and also to test basic theoretical propositions in the developmental and learning sciences.

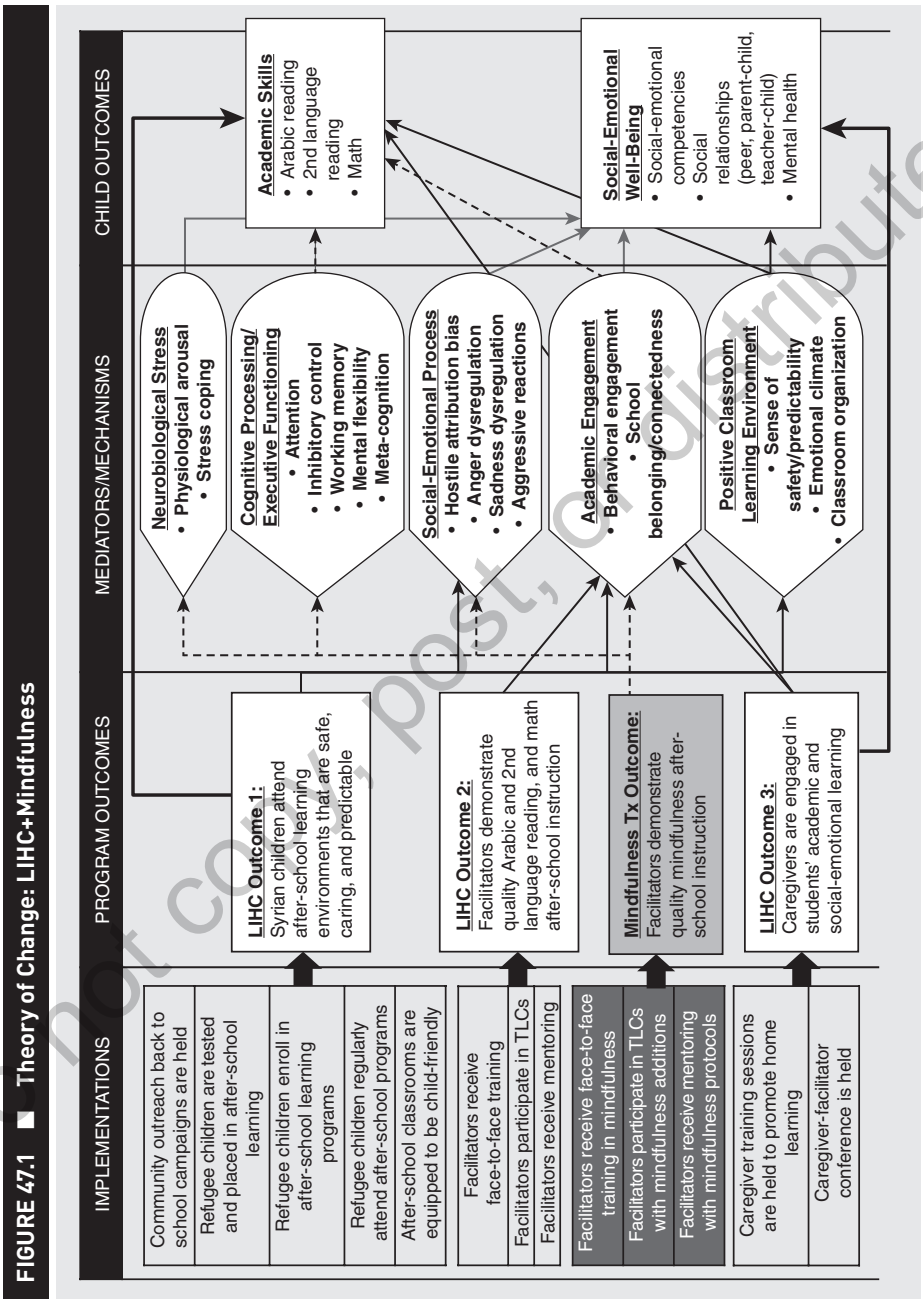
For the last seven years, I have been collaborating with the International Rescue Committee (IRC) to rigorously evaluate one of their interventions called "Learning in a Healing Classroom" (LIHC; Aber et al., 2017a, 2017b). We developed a theory of change that describes how the intervention is supposed to change features of children's social-emotional processes and in turn their literacy and numeracy skills and their behavior problems. (See Figure 47.1.) As you see, we hypothesized that LIHC would improve four features of their social-emotional development: hostile attribution bias (HAB: the tendency to attribute hostile intent to an ambiguous behavior of a peer), anger dysregulation (AD: the tendency to become angry in a social situation), sadness dysregulation (SD: the tendency to become sad in a social situation), and aggressive reactions (AR; the tendency to behave aggressively in a social situation).

The measures used to index these constructs entail the use of stories (hypothetical vignettes) that ask children to imagine and report on how they would think, feel, and behave in challenging situations with peers at school. Such measures have been used successfully in intervention research for several decades in the United States (Dodge et al., 2015; Aber, Jones, Brown, Chaudry, & Samples, 1998). Because "self-regulation" is considered by many researchers and practitioners to be a cross-culturally important construct, recent studies have begun to test the reliability and validity of its measurement across quite different Western cultures (Dodge & Frame, 1982; Dodge et al., 2015; Di Giunta et al., 2017). The results of these efforts to establish the reliability and validity of these measures of HAB, AD, SD and AR were so promising that we decided to use modest adaptations to them in our impact evaluations of LIHC in the African and Middle Eastern cultures of Niger and Lebanon and to check the reliability and validity of the measures in these non-Western cultures using baseline (preintervention) data collection with the participating kids. This is where the mistake arose.

We used responses to six stories to measure each of the four constructs' "reliability." In this case, reliability meant that children's responses on each story were positively correlated with children's responses on the other five stories. The most common way to describe this type of reliability is with a summary statistic called "Cronbach's alpha." Alphas range from zero (there is no correlation among responses to the items in the scale) to 1 (there is perfect correlation among responses). Generally, researchers trust measures as reliable if alpha ≥ 0.70.

In our study in Lebanon, alpha ranged from 0.79 to 0.89 in our sample of 5- to 16-year-old kids. This meant that children's responses to different stories were highly correlated with each other, as they should be if they were measuring HAB, AD, SD, and AR. Also, all four measures were positively correlated with each of the other measures, as prior theory and research in Western countries would predict. I found these preliminary analyses so encouraging that I concluded that the measures of HAB, AD, SD, and AR were not only reliable but cross-culturally valid.

FIGURE 47.1 ■ Theory of Change: LIHC+Mindfulness



I failed to notice that there were severe "floor effects." What this means is that the vast majority of children responded "not at all likely" to nearly all of the items meant to measure anger dysregulation, sadness dysregulation, and aggressive reactions. This is a pattern of response that is not typically seen in Western samples of children.

It was only after testing the impact of LIHC on children's measures of self-regulation at "midline" (after half a year of intervention) that I began to question the cultural validity of the reliable and positively correlated measures. The intervention reduced children's hostile attribution bias (as we predicted), but it increased children's anger dysregulation and sadness dysregulation (against our predictions)! How could that be if these were reliable and valid measures of these social-emotional processes for refugee children in Niger and Lebanon?

We are left with several different ways to understand these puzzling findings. First, perhaps LIHC, despite its intention and design, actually made anger dysregulation and sadness dysregulation worse among Syrian refugee children in Lebanon. This would be a most serious problem, because it would be a violation of the Hippocratic oath: first do no harm! Second, perhaps the theory that low AD and SD are indices of positive adaptation doesn't apply to Syrian refugee children in Lebanon. It may be that expressing some anger and sadness dysregulation is more adaptive than expressing none, especially in conflict-affected countries. Third, perhaps these measures of AD and SD simply are not cross-culturally valid among Syrian refugee children in Lebanon. Perhaps anger and sadness dysregulation measured in this way is imposing Western definitions of these concepts on a complex Middle Eastern culture that views emotion expression in very different ways.

In short, are the puzzling findings due to the program, the theory, or the measures? In relatively low-stakes research, using a measure of uncertain validity is a small and reparable mistake. But in relatively high-stakes research, such as field experiments of interventions designed to serve vulnerable children, using measures of uncertain validity in a new culture is a big and much less reparable mistake. That's because practical actions (like policy and funding decisions on whether to continue to invest in interventions like LIHC) hinge precariously on results from such experiments. By using several measures of uncertain cross-cultural validity, I made a research mistake with serious potential implications both for the developmental and learning sciences and for the real world of programs for refugee children.

I believe I made this mistake because I judged that the reliability of the measures constituted adequate evidence of their cross-cultural validity. From this experience, I relearned how important it is to use measures of clear validity within new cultures. The best way to avoid this mistake is to use measures of key constructs that are already well validated within the culture in which the study is being conducted. But this option was not available to us at the time due to funding and time constraints. We faced a tough choice: We could use measures validated in other cultural contexts, or we could fail to measure key constructs in the theory of change. I consider using measures not yet cross-culturally validated to be a real mistake. But I consider not trying to measure concepts critical to the evaluation to be a bigger mistake.

REFERENCES

Aber. L., Brown, J., Jones, S., & Roderick, T. (2010). SEL: The history of a research–practice partnership. *Better: Evidence-based Education, 2*(2), 14–15.

Aber, L., Brown, J. L, & Jones, S. M., Berg, J., & Torrente, C. (2011). School-based strategies to prevent violence, trauma and psychopathology: The challenges of going to scale. *Development and Psychopathology, 23*(2011), 411–421.

Aber, J. L., Jones, S. M., Brown, J. L., Chaudry, N., & Samples, F., (1998). Resolving conflict creatively: Evaluating the developmental effects of a school-based violence prevention program in neighborhood and classroom context. *Development and Psychopathology, 10*(2), 187–213.

Aber, J. L., Torrente, C., Starkey, L., Johnston, B., Seidman, E., Halpin, P., . . . Wolf, S. (2017a). Impacts after one year of "Healing Classroom" on children's reading and math skills in DRC: Results from a cluster randomized trial. *Journal of Research on Educational Effectiveness,10(*3), 507–509.

Aber, J. L., Tubbs, C., Torrente, C., Halpin, P. F., Johnston, B., Starkey, L., . . . Wolf, S. (2017b). promoting children's learning and development in conflict-affected countries: Testing change process in the Democratic Republic of the Congo. *Development and Psychopathology, 29,* 53–67.

Ainsworth, M., & Bell, S. (1970). Attachment, exploration and separation: Illustrated by the behavior of one-year-olds in a strange situation. *Child Development, 41*(1), 49–67.

Di Giunta, L., Iselin, A. R., Eisenberg, N., Pastorelli, C., Gerbino, M., Lansford, J. E., . . . Thartori, E. (2017). Measurement invariance and convergent validity of anger and sadness self-regulation among youth from six cultural groups. *Assessment 24*(4), 484–502.

Dodge, K. A., & Frame, C. (1982). Social cognitive bias and deficits in aggressive boys. *Child Development, 53(3),* 620–635.

Dodge, K. A., Malone, P. S., Lansford, J. E., Sorbring, E., Skinner, A. T., Tapanya, S., . . . Pastorelli, C. (2015). Hostile attributional bias and aggressive behavior in global context. *Proceedings of the National Academy of Sciences, 11*2, 9310–9315. doi:10.1073/pnas.1418572112

Jones, S. M., Brown, J. L., & Aber, J. L. (2011). Two-year impacts of a universal school-based social-emotional and literacy intervention: An experiment in translational developmental research. *Child Development, 82(*2), 533–554.

Mischel, W., Shoda, Y., & Rodriquez, M. L. (1988). Delay of gratification in children. *Science, 244*(4909), 933–938.

CRITICAL THINKING QUESTIONS

1. Do you think that children's ability to regulate their angry and sad emotions are critical abilities in countries as different as the United States, Lebanon, and Niger? Why, or why not?
2. Do you think that researchers can reliably and validly measure anger and sadness dysregulation using the same methods in very different cross-cultural contexts? Why, or why not?
3. How serious a mistake do you think this researcher made?
4. Is there anything you could recommend to the researchers to help avoid making this mistake in the future?