Promises and Perils of Assessing Character and Social and Emotional Learning

Clark McKown, Ph.D.

Paper presented at the National Academies of Sciences, Engineering, and Medicine Workshop
on Approaches to the Development of Character

July 27, 2016

Author Note

Clark McKown, Department of Behavioral Sciences, Rush University Medical Center.

This work was supported by Institute of Education Sciences through Grants R305A110143 and R305A140562 to Rush University Medical Center. The opinions expressed are those of the author and do not represent views of the Institute or the U.S. Department of Education.

Correspondence concerning this article should be addressed to Clark McKown, Department of Behavioral Sciences, Rush University Medical Center, RNBC 4711 Golf Road, Suite 1100, Skokie IL 60076. E-mail: Clark_A_McKown@rush.edu

A friend and colleague once said, “What gets assessed gets addressed” (R. Weissberg, *personal communication*). And here is a corollary: If we want character and social and emotional learning skills to be addressed in schools and youth development program, we had better assess those skills. Dr. Card’s paper provides useful and clear standards for determining what good measurement is in terms of reliability, validity, and measurement equivalence. Dr. Card points out important problems in the science of character education, such as when a construct is measured differently in every study, or when it is measured the same way in every study. Dr. Card provides important guidance to the field about the planning, execution, and dissemination of research on measurement. His paper will surely help to shape measurement practices in the scientific study of character education and social and emotional learning.

Here I raise four points that build on Dr. Card’s work as it relates to the field of character development and social and emotional learning. First, advancing assessment in the field requires a vigorous pursuit of conceptual clarity. Second, the field will benefit from efforts specifically to create assessments designed for practice, and those efforts should include consideration of how assessment data are interpreted and used. Third, I highlight the importance in practice of being clear about the purposes for assessing character and social and emotional learning. And finally, I argue that the method of assessment is a critical but underappreciated consideration, because different methods of assessment are suited to measuring different dimensions of character and social and emotional learning.

Conceptual Clarity in a World of Fuzzy Boundaries

In his paper, Dr. Card speaks of “fuzzy boundaries,” a reference to the difficulty of defining the conceptual line between constructs. Dr. Card uses that term to refer to the fuzzy boundaries between one construct and another. However, the metaphor is relevant to the entire field and conceptual fuzziness is a prominent theme of this conference. A thought experiment will illustrate how: Today, if ten scientists and practitioners were asked to define character or social and emotional learning, it is highly likely that ten distinct definitions would emerge. They would likely overlap, but I fear they would be more like be an archipelago of fuzzily-bounded conceptual islets. So the broader fuzzy boundary problem is that there is no consensus about what constitutes character or social and emotional learning.

This is consequential. It is the origin of what I see as a kind of measurement paralysis wherein there are not many robust measurement development efforts because everyone is waiting for clarity before committing the considerable resources needed to build sound measurement systems. Paradoxically, fuzzy boundaries also levy an implicit tax on the field. Without clarity, one can spend a lot of resources purchasing, creating, or adapting measures and taking up precious time, and end up with nothing. Some might argue that in this imperfect world of social science and its fascinating subjects, fuzziness is inevitable. In general, I would agree. But greater clarity in the field is possible and essential for its healthy forward momentum.

Let me share an example of what conceptual clarity makes possible. My colleagues and I have been working on building scalable, web-based systems to measure social-emotional skills in the elementary grades (McKown, Russo-Ponsaran, Allen, Johnson, & Russo, 2016). We divide those skills into specific thinking skills, like the ability to understand another person's thoughts and feelings and solve social problems, and behavioral skills, like the ability to join an ongoing group and help someone in need. We also include self-control, which has both mental and behavioral components. In addition, we believe that peer social networks are powerful indicators and outcomes of social-emotional skill.

In our effort to capture what is most important, we have made commitments about what is and what is not included in the social and emotional arena, which has given us the clarity of purpose needed to build robust measurement systems that largely meet the standards articulated by Dr. Card. Our commitment to a model has very specifically and strongly influenced assessment design considerations. We do not claim to have a perfect answer. However, it is surely a good sign that colleagues from different “camps” who care about children’s social and emotional development have asked us to partner with them to provide measurement and assessment support. I urge scientists and practitioners alike to be diligent about clarifying precisely what is being measured. This will stimulate the development, adaptation, and adoption of assessments that are sound and useful.

Practical Assessment and Its Consequences

Dr. Card’s paper focuses largely on assessment for science. There is also an urgent need for good assessments for use in practice. In addition to the aspects of validity that Dr. Card described, for practice, assessments should demonstrate what Samuel Messick called “consequential validity,” which refers to the ways in which test scores are interpreted as a basis for action, and the consequences, both intended and unintended, of those actions (Messick, 1995). If this sounds esoteric, a real-life example will show that it is not. The CORE districts is a consortium of ten large school districts in California who have been using self-report measures of self-efficacy, social awareness, mindsets, and self-management as part of their accountability system. I believe what they are doing is a bold and important experiment—using measures of these skills to determine how well schools are doing their jobs. But not many months ago, a very public controversy unfolded on the pages of the New York Times, with prominent figures in the field criticizing this endeavor, arguing that the measures did not have the qualities that justified their use for accountability (Duckworth, 2016).

At issue in the CORE districts was consequential validity, with the key question being this: Are the measures of character chosen by the CORE districts appropriate indicators of school performance and are the scores they yield a reasonable basis for accountability-related consequences? Reasonable people can and have debated the answer to this question. That is not the point. The point is that the consequential validity of all measures of character and social and emotional learning (and, by the way, achievement) is a crucial consideration. For any measure, consequential validity can be only partly evaluated by rigorous study of the measure's technical properties. At least some of a measure's consequential validity is a matter of social values. In considering the validity of measures, if we are being complete in our work, we cannot therefore be totally insulated from the vicissitudes of social values and our historical moment in its glorious complexity.

From its inception, our work has considered the meaning and consequences of the measure we are developing. And yet like our friends at CORE, we do not claim we have it all worked out. We know how to present data on children's social and emotional understanding and social relationships so that teachers understand. We know how to talk with teachers about what assessment data mean. We know and can describe some of the effective actions educators can take to address the issues that the assessment surfaces. However, it remains largely untested whether educators will (or should) interpret and use our assessment's data as intended. And it is even less clear what happens if they interpret and use the assessment data in ways we never imagined. Like those at the CORE district, we take the position that any worthwhile assessment effort will require experimentation and revision, and some tolerance for side effects.

Sense of Purpose in Assessment

A clear intention can place constructive boundaries on the consequences of assessment. Contrast fictitious Programs A and B. Leaders of Program A have decided to measure many dimensions of character and determine the use of those measures afterwards. In contrast, leaders of Program B have decided to measure particular social and emotional skills specifically and exclusively formatively to inform instructional planning. In Program A, how assessment data will be interpreted and used is unclear. Therefore, the possibility that it will be used for non-valid and potentially harmful purposes is high. In addition, in Program A, because no one is clear about the goals and therefore payoff of assessment, it is likely that considerable resources will be expended on assessment that will not yield any benefit.

In contrast, in Program B, all players know the purpose of assessment. Therefore, they will expect the data to be used in a particular way, increasing the likelihood that it will be used. In Program B, because the purpose of assessment is clear, training in the interpretation and use of assessment data can be focused and practical. This will increase the odds the data will be used appropriately. In Program B, all player understand a large number of decisions that will *not* be informed by the data—school and teacher accountability, special education placement, etc. Therefore, after data are collected, constituents will be less anxious that data may be used against them. It is still of course possible that in Program B, formative assessment data will have negative unintended consequences, but the range of those negative consequences has been significantly reduced.

As practitioners consider implementing assessments, it is important to note that at the present moment, the purposes for which social and emotional assessment can be fully used are limited. Good character and social and emotional learning assessments, including ours, can help clarify student need and be used to inform instruction. In other words, the current state of the art supports, in my opinion, formative assessment. In addition, existing assessments are promising for program evaluation. However, few character and social and emotional learning assessments can be used for high-stakes accountability purposes because none have been developed that have the psychometric properties, consequential validity included, to support this use.

Wisely Selecting Methods of Assessment

Finally, there is a rarely-considered but critical consideration in the assessment of character and social and emotional learning. Specifically, in the best of all worlds, the method of assessment should be matched to what is being measured. By method of assessment, I am referring formally to the procedure through which an assessment samples behaviors hypothesized to reflect an underlying character or social and emotional learning skill. In discussions of assessment, surveys are often given as examples—in a survey, respondents rate items, often indicating how true of the respondent are a variety of self-statements. However, there are many other methods of measurement, including observation, teacher ratings, a variant called direct behavior ratings (http://dbr.education.uconn.edu/) , and direct assessments, in which children demonstrate their skill through solving challenging problems (McKown et al., 2016).

Here is the important part: no single method can measure everything well and each method is better suited to measuring some things than others. To assess how well a child reads, we can ask her to fill out a self-report questionnaire in which she rates her reading skills. But a sound direct assessment of reading—in which she reads something and answers questions about what she read, for example—is likely to prove more informative. Similarly, to measure at how well a child reads facial expressions, we can ask children to rate his skill level. But I would venture to say that direct assessment, in which he looks at faces and indicate what emotion each face is expressing, will be far more valid. To measure behavior, teacher report is probably better than self-report, and vastly more practical than observation. To measure peer acceptance and networks, peer nominations are superior to teacher report and other methods. Reasonable people can disagree about what method is best-suited to measuring what construct. What is important is that researchers and practitioners seriously consider what method of assessment is best for what they want to assess. A good general operating assumption is that the best assessment of character and social and emotional learning employs multiple methods and multiple raters.

The Stakes

The stakes are high. Yes, the scientific study of character education, which is the focus of Dr. Card's paper, depends heavily on developing some consensus about what to measure and how to measure it. But it is much bigger than that: No less than the survival of the character education and social and emotional learning enterprises—from policy to practice to research—depends on our ability to assess these skills well and rigorously. How else can we know what children's strengths and needs are and therefore how to target instruction to foster character? That is formative assessment. How else can we know if a set of practices intended to foster character worked? That is program evaluation. How else can we know to what heights of character development students have risen? That is perhaps summative assessment. How else can we know if our system of education has met relevant state standards?

These are not idle questions. If nature abhors vacuums, educational fads thrive on them. Without evidence, rooted in good measurement, the pendulum tends to swing from one fad to another. All of us—scientists, practitioners, and policy makers alike—should hope that the very best evidence of what works will be used to spur the evolution of effective youth development programs and practices. Good measurement is foundational to collecting such evidence. Without good evidence, rooted in rigorous measurement, there is great risk. If we do not measure character and social and emotional learning skill well, these fields will be buffeted by the winds of fad and polemics and they risk ending up on the dust-pile of bygone movements.

Summary

In summary, in addition to Dr. Card's thoughtful and useful recommendations, there are four important considerations: getting to conceptual clarity; designing assessment for practice; being clear about the purposes of assessment; and selecting the method of assessment best-suited to what it is we want to measure. These questions are important enough that many of the people at this conference, under the leadership of Roger Weissberg and Jeremy Taylor from CASEL, are working to advance the field of social and emotional assessment. The stakes are high. Let us move swiftly and wisely to build practical, useful and scientifically sound measures of character and social and emotional learning.

References

Duckworth, A. (2016, March 26). Don't grade schools on grit. *The New York Times.*

McKown, C., Russo-Ponsaran, N.M., Allen, A.A., Johnson, J., & Russo, J. (2016). Web-based direct assessment of children's social-emotional comprehension. *Journal of Psychoeducational Assessment,* 34, 322-338. DOI: 10.1177/0734282915604564

Messick, S. (1995). Validity of psychological assessments: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American Psychologist,* 9, 741-749, doi: 0003-066X/95