Journal of Youth Development
Bridging Research & Practice

The Choices, Challenges, and Lessons Learned from a Multi-Method Social-Emotional / Character Assessment in and Out of School Time Setting

Valerie B. Shapiro
University of California, Berkeley
VShapiro@Berkeley.edu

Sarah Accomazzo
Center for Prevention Research in Social Welfare

Jennette Claassen
Playworks, Inc.

Jennifer L. Fleming Robitaille
Devereux Center for Resilient Children

Author Notes: This project was supported by the Kaiser Permanente Northern California Community Benefit Program. The authors wish to acknowledge all of the Playworks staff who participated in this project, and specifically the coaches, program managers, program associates, and program directors from the Silicon Valley and San Francisco offices. The authors also wish to thank Dr. Nicole Russo and Jason Johnson (Rush Neurobehavioral Center), J.T. Folsom-Kovarik (SoarTech), and Jessica Adamson (Apperson Evo) who were each generous, creative, and patient collaborators playing an essential role in the completion of this work. Although one author of this paper had a role in the development of the Devereux Student Strengths Assessment (DESSA), no author receives any financial remuneration from the sale of the DESSA, VESIP, or any other tool or resource mentioned within this manuscript.

JOURNAL OF YOUTH DEVELOPMENT
bridging research and practice

Volume 10, Number 3, Special Issue 2015 **Article 151003FA003**

The Choices, Challenges, and Lessons Learned from a Multi-Method Social-Emotional / Character Assessment in and Out of School Time Setting

Valerie B. Shapiro
University of California, Berkeley

Sarah Accomazzo
Center for Prevention Research in Social Welfare

Jennette Claassen
Playworks, Inc.

Jennifer L. Fleming Robitaille
Devereux Center for Resilient Children

Abstract: Out-of-School-Time (OST) programs are increasingly recognized as a venue to actively engage children and youth in character development activities, but little guidance exists as to how to assess individual children and youth in OST environments for the sake of evaluating their character development. This research brief uses an illustrative case study to reflect upon the experience of selecting and completing a strength-based, multi-modal social-emotional / character assessment that used a direct assessment and a multiple informant behavior rating scale in an OST setting. Insights derived from the case study reveal opportunities and challenges associated with each assessment modality. This paper shares lessons learned with those conducting individual assessments in OST environments and with those seeking to improve our capacity to complete screening, formative, and summative assessments of social-emotional and character constructs in OST youth development programs to help children thrive.

Introduction

Positive Youth Development (PYD) programming integrates prevention and intervention strategies to develop safe and healthy behaviors, skillful and respectful social interactions, positive engagement with family and community, and habits and dispositions for school and life success (National Clearinghouse on Families & Youth, 2007; Weissberg, Durlak, Domitrovich, & Gullotta, 2015). The consensus ideals of PYD programming include the promotion of character development (Catalano, Toumbourou, & Hawkins, 2014).

Out-of-school time (OST) settings are increasingly recognized as settings in which character development can be promoted. Over 10 million children in the United States participate in OST programming after school, engaging nearly 25% of American families (Afterschool Alliance, 2014). OST programs, however, vary widely (Gullotta, 2015). At a minimum, OST programs provide children with a safe and supervised environment when children are not engaged in teacher-guided instruction during the school day and outside of school hours. Some program models provide support for children completing their homework or resources for free-play. Other programs may offer enriching activities and experiences intentionally designed to promote youth development. OST programs may implicitly influence the development of youth character, as do all settings in which youth spend time, or they may explicitly focus on the development of youth character, and thus be considered among the providers of Character Education.

Character is the "composite of psychological characteristics that impact the child's capacity and tendency to be … socially and personally responsible, ethical, and self-managed" (Berkowitz & Bier, 2005, p.2). Thus, character education is "any deliberate approach" to teach "children about basic human values, including honesty, kindness, generosity, courage, freedom, equity, and respect" with the goal of raising "children to become morally responsible, self-disciplined citizens" (Association for Supervision and Curriculum Development, n.d.). Character education is a popular idea, but its ultimate success depends upon conducting evaluations of its effectiveness and using strategies that have been tested and demonstrated to be effective in achieving the objectives of character education (Leming, 1993).

This special issue of the *Journal of Youth Development* aims to bring together empirical findings to understand key questions about the effectiveness of character education in OST environments, including a) the facets of character development promoted by specific features of youth development programs, b) the effect of program designs on character development in youth, and c) determining for which youth, during which periods of development, different facets of character development programming are most impactful. These are important questions for advancing the science and practice of character education in OST environments. Yet, we assert that our ability to answer these questions is entirely dependent on our capacity to measure the growth of youth character in OST environments.

This paper is jointly authored by the Director of Evaluation of a non-profit provider of OST programming that serves half a million children annually and the three researchers she consulted when trying to devise a measurement strategy to assess character development within her organization. We have thus prepared a case study and articulated resulting "lessons learned" to fill a gap in the current literature for how to practically and routinely measure youth development constructs of character in OST settings, as a requisite for answering important questions about the effectiveness of character education as implemented in OST environments.

The Need for an Assessment

Playworks (www.Playworks.org) is a national organization founded by Jill Vialet in 1996. Program developers designed Playworks to yield many positive youth developmental outcomes, including character. Full-time, trained Playworks staff provide multi-faceted, play-based OST programming on-site at low-income schools through five distinct services. First, Playworks provides traditional OST programming that emphasizes scholastic support, physical activities, and group projects outside of school hours. Second, Playworks staff teach weekly classes during the school day where students participate in inclusive, cooperative physical activities to learn basic sports, games, and lessons on physical health, fitness, violence prevention, and safety. Third, Playworks staff structure daily recess to reinforce and promote healthy, inclusive, skill-building, and engaging play. Fourth, Playworks staff provide leadership training through a "Junior Coach" program, where small teams of students work together to learn games, principles of fair and inclusive play, and positive conflict resolution in order to teach these skills and lessons to their peers and younger classmates as recess leaders. Finally, Playworks uses interscholastic and developmental sports leagues to provide opportunities that students may not otherwise have.

Evaluations of Playworks programming have demonstrated many positive outcomes, including improvements in school climate (London, Westrich, Stokes-Guinan, & McLaughlin, 2015), reductions in bullying (Fortson, James-Burdumy, Bleeker, Beyler, London, Westrich, Stokes-Guinan, & Castrechini, 2013), and improvements in learning/academic performance (Fortson et al., 2013). However, despite anecdotal reports from coaches, teachers, principals, and parents about the impact of Playworks on character development, there has been limited empirical evidence thus far on of the effects of Playworks on youth character. We suspect this is largely due to our limited capacity to efficiently and effectively measure character constructs. Thus, the question remains as to whether Playworks, like many programs purporting to promote character, is actually accomplishing its character-oriented objectives. Of the 41 programs reviewed by the U.S. Department of Education's Institute of Educational Sciences What Works Clearinghouse in 2008, only seven were determined to have adequate evidence of programmatic impact on moral and character attributes (Berkowitz, Battistich, & Bier, 2008).

The Character Education Partnership also published a review, including character promoting *practices* in addition to *programs*, and using a more inclusive definition of what constitutes "scientific support," and found 33 effective character promotion strategies (Berkowitz & Brier, 2005). Both of these reviews restricted their scope to programs that occurred during the regular school day, however, limiting the potential of character promoting OST programs to reflect and learn from these syntheses.

Proponents of character education debate which, if any, of these reviews employed a useful approach for summarizing the state of effectiveness research in character education. Even in the current era of evidence-based policy-making and practice, none of these reviews has yet to have a notable impact on the evolution of practice in character education. Leming (2008) noted that three of the most widely used character education programs in the United States, collectively claiming to impact over 32 million youth, are not included in any of these reports. Given the tensions between research and practice that have shaped the field of character education for decades, Leming suggests using an engineering-based process rather than a science-based process for research, suggesting that this would enable greater alignment between the alleged producers and consumers of knowledge. This approach would have research be "less focused on developing generalizable views on how schools and pedagogy work and would instead be more directly concerned with the development of high quality solutions to practical problems." (p. 151) Quoting from Shaver (2001, p.233), he further explains, "Engineering is technology, not science, not even applied science. It is a different type of research enterprise with a different epistemology. The purpose of engineering is (not to create more

knowledge) practical and set in a social context. The purpose is to create artifacts that serve humans in a direct and immediate way. Knowledge is generated and used in the design, production, and operation of artifacts that meet recognized social needs."

We believe that character education would benefit from a research enterprise that intersected both science and engineering to offer practitioners tools to critically examine their own work and continue evolving strategies toward greater effectiveness. We believe this integrative approach begins with putting practical assessment tools into the hands of practitioners to enable learning from data in real time within routine practice. This ideal is already incorporated in the 11 Principles of Effective Character Education published by Character.org (Sipos & Maupin, 2010). The 11th principle calls for the regular assessment of "culture and climate, the functioning of its staff as character educators, and the extent to which its students manifest good character." (p.22). Specifically, character programs are directed to assess "student progress in developing an understanding of a commitment to good character and the degree to which students act upon the core values." (p.23). This is exactly the type of assessment that Playworks sought.

To uncover state-of-the-art (and perhaps state-of-the-science or engineering) tools for the assessment of Character Education, the *Primer for Evaluating a Character Education Initiative* (Berkowitz, 1998) was consulted. Recommendations within are directed toward practitioners who would like to appraise and evolve their own practice in the most rigorous ways possible. This resource suggests that practitioners should use high quality existing instruments to appraise their work. They imply that existing resources will have been tested for validity (that they actually measure what they claim to be measuring), reliability (that they will do so consistently across uses), and that they will offer comparison data against which one can interpret their own data. They also assert that practitioners should attend to other important features of the tool, such as the applicability, length, required materials and training, administration format, and cost. The authors acknowledge that one of the most frequent questions that they receive is where/how to locate and appraise such tools. They refer readers to a list of options collected by the Character Education Partnership and placed on their website that was under development at the time of the *Primer*'s publication.

Thus, we went to the Character.org website and looked at the impressive collection of potential assessment tools that could be used to measure the growth of individual character. At the time of this writing, 72 tools were listed, but little guidance was available as to whether the tools were practical or scientifically-sound (in accordance to the aforementioned criteria) and none explicitly referenced utility of the tool for OST settings. Thus, we consulted the newly released *Handbook of Moral and Character Education* (2014) in hopes of finding a framework or synthesis to inform the selection of a tool to pilot. This volume, however, did not include a chapter on Assessment, nor did it even list the term "assessment" "evaluation" or "outcome" in its Subject index.

On the other hand, the *Handbook of Moral and Character Education* does highlight the interdependence of two related character concepts: ethical character and performance character (Elias, Kranzler, Parker, Kash, & Weissberg, 2014). This work suggests that children can have a highly developed sense of morality (ethical character), but be unable to enact their convictions (performance character). Alternatively, children can be skilled actors (performance character), regulating and manipulating themselves and their social environments, but may not have their activity steered by a clear moral compass (ethical character, Character Education Partnership, 2008). The practice of Character Education has historically emphasized the development of ethical character while a related field, Social Emotional Learning (SEL), has historically emphasized the development of what may be regarded as performance character. SEL aims to help students "better manage their own emotional state and their interactions with other people" (Elias et al., 2014, p.278).

Accepting the argument made by Elias et al. (2014) that the related fields of SEL and Character Education are interconnected, complementary perspectives that stand to benefit a common research, practice, and advocacy agenda, the newly released *Handbook of Social Emotional Learning* (2015) was also consulted for assessment guidance. This handbook is considered to be the comprehensive and definitive volume on SEL research, practice, and policy, and emerges from a tradition that has had a greater emphasis on standardized assessment over time. We found that a major section of the SEL *Handbook* was devoted to the topic of assessment. The volume conveys that the function of SEL assessment includes: (a) screening to determine which students should receive programming and at what intensity; (b) formative assessment to identify gaps or needs which could inform individual or programmatic goals, the selection of strategies, and monitoring progress, and (c) summative assessment to evaluate success relative to baseline or against a standard (Denham, 2015). These purposes seemed largely applicable to the needs of character educators. Chapters within the section on assessment are devoted to the assessment of environments (the organizations and organizational agents shaping the delivery of SEL programming) and of individuals (social-emotional comprehension and competence in children and youth). The four chapters focused on individual assessment (our present focus) examine the assessment of SEL in children and youth in school environments during school time, but still leave much to be imagined about the assessment of SEL in children and youth in school environments during OST.

Methods of Collecting Data Related to SEL/Character

There are many methods of collecting information about the capacities of children and youth. The most popular methods include direct observations and behavior rating scales (Elliott, Freu, & Davies, 2015; McKown, 2015). Although some argue that direct observation is a superior method for collecting valid information, it is best suited for observing the frequency of a small number of discrete behaviors within a single domain. Research indicates that a minimum of five 30-minute observations are required to reliably count instances of even high-frequency, readily observable behavior (Doll & Elliott, 1994). In addition to substantial observation time, the time required for training, establishing coder reliability, and scoring poses considerable feasibility barriers in routine practice (Denham, 2015). This method can also require each program to determine standards or criteria for defining, interpreting, and making decisions based upon the observed data, rather than contextualizing the behavior relative to a representative, national norm, and empirically determined practice guidelines (Naglieri et al., 2013).

Behavior rating scales also strive to ascertain the frequency of behavior, but do so through a series of questions posed to an informant, retrospectively offering a general impression of the frequency of behavior. Behavior rating scales typically have broader coverage than direct observation protocols. Although behavior rating scales are faulted for potential subjectivity and recall bias, efficiency demands have made them the most prevalent assessment method (Elliott et al, 2015). Relative to direct observation, they may be a preferable approach for assessing low-frequency, high-impact (memorable) behavior. Behavior rating scales can be used repeatedly, across settings, and with multiple informants to capture a comprehensive understanding of the child's behavior, over time, and relative to a standardized reference group. When the same child is assessed by different informants who experience the child in different environments, however, scores tend to converge only moderately, leading to little consensus on the "true" level of the child's current capacities. Behavior rating scales can use self-assessment to gain insight into less observable behaviors. Although highly valuable information, this approach can be even more subjective and present a formidable challenge with young children.

Both direct observation and behavior rating scale methods can be used to assess the execution of behaviors related to any construct, such as social-emotional competence or character (McKown, 2015). Because of their differences in coverage, they are unlikely to be directly comparable. Studies comparing direct observations to behavior rating scales have found correlations ranging from r=.05 to r=.52 (Merrell, 1993).

Two primary conclusions can be drawn about the assessment of character performance in youth from the *Handbook of Social Emotional Learning*. First, because there are distinct advantages of different assessment modalities, "multi-method, multi-rater assessment is preferred over mono-method, mono-rater assessment." (McKown, 2015, p.332). Second, similar to the guidance provided in the *Primer for Evaluating a Character Education Initiative,* criteria for selecting tools should include the adequacy of user documentation, strength of psychometric properties, relevance to the populations of students to be assessed, practicality of administration, inclusivity of multiple perspectives and dimensions, and potential for interpretable and meaningful information to guide decisions (Denham, 2015).

Recall that the recommendations across these resources, however, have been primarily written for the use of administrators, teachers, and student support personnel in academic school-day contexts. Little consideration is given to the ways in which each assessment modality may be differentially suited for OST programs relative to these criteria. Compared to classrooms, OST settings may face higher student to staff ratios, lower levels of engagement of eligible informants or qualified assessment administrators, less support infrastructure or staff time without student responsibilities, less staff training or versatility with data collection and interpretation, less access to technology, and smaller budgets for evaluation activities. Furthermore, there may be more varied activities and service environments between and within programs, briefer program intervals (e.g., 6-week summer sessions), and more commingling of children of varying ages, and ability levels. Finally, staff and student turnover (as well as irregular student attendance) is often far higher in OST settings than in classroom settings.

Yet, as OST settings become increasingly viewed as opportunities to foster character development, assessment is needed to guide and monitor this work. Many funders, host agencies, and regulation bodies are demanding the use of assessment for screening, formative, and summative purposes. Accordingly, the National Afterschool Association (2011) recently asserted that PYD professionals need competence in tasks such as 1) observing and assessing individual needs, 2) describing and understanding the program participants, 3) showing whether children are benefitting from the program, and 4) using assessment information to modify or enhance existing program activities. The case study that follows emerged from one specific attempt to conduct a multi-modal assessment of children's SEL/character competencies in an OST setting.

The Current Study

This research brief presents a case study whereas the logic espoused within these various resources was applied to the actual assessment of SEL and character constructs among individual students in OST settings. We present this case study to share the experience of conducting assessments in OST settings in order to illuminate lessons learned and encourage further engineering of useful tools for this purpose.

Methods

Participants

Participants in Playworks programming reflect the demographic make-up of the lower income communities in which Playworks programming occurs. An assessment pilot, using a behavior rating

scale and a direct observation protocol, were largely undertaken in the Northern California region, where the gender make-up is roughly even, and assessments indicate that students identify as approximately 12% Asian/Asian-American, 12% Black/African-American, 43% Hispanic/Latino(a), 10% White/European-American, and 23% "Other / Multiple-Race."

Selecting a Behavior Rating Scale

The *Handbook of Social Emotional Learning* was relied upon to select a behavior rating scale that could be used with elementary school children. This *Handbook* favorably reviews the *Behavioral and Emotional Rating Scale (BERS-2*; Epstein, 2004), *Devereux Student Strengths Assessment* (*DESSA*; LeBuffe, Shapiro, & Naglieri, 2009/2014), *Social and Emotional Assets and Resilience Scales* (*SEARS*; Merrell, 2011), and *Social Skills Improvement System - Rating Scale* (*SSIS-RS*; Gresham & Elliott, 2008) against previously stated criterion.

Additional criteria were applied to consider the utility of these tools in OST settings. Some OST settings require administration and interpretation by personnel without a graduate degree. Some require that assessments can be done based on limited observation time, since many programs are themselves short and/or give adults limited exposure to any one child. Finally, OST settings might prefer a single form and scoring procedure for children of diverse ages and ability levels, since there is more likely to be heterogeneity among the students than may be experienced in a typical school setting.

Each of these high-quality existing instruments (*BERS-2*, *DESSA*, *SEARS*, and the *SSIS-RS*) are classified as Level B assessments on their respective websites, which implies that their use can be overseen by somebody with some graduate coursework in assessment, but does not require that person to hold an advanced degree in psychology or a related field. Relative to other assessments these tools may be particularly low risk for use by personnel with less training, because they are entirely strength-based and thus may have less potential for stigmatizing or labeling individual children than those that are used to indicate pathology or identify problem behaviors.

There are some ways, however, in which the tools differ. Although some of these assessments require "regular, daily contact with the child or adolescent for at least a few months before responding to the rating scale" (Epstein, 2004, p.12), the *DESSA* only requires raters to have contact with the child for an average of six hours a week over a four week period. Most of these assessments have different forms or scoring procedures (norms) for children of different ages or abilities. The *DESSA* utilizes one rating form and norms table for all children and youth in Kindergarten through 8th grade, regardless of individual child characteristics. In addition, the *DESSA* is the only one among these four tools that presents studies in the test manual that explore appropriate use with OST staff, and provides norms explicitly applicable to these raters.

Given the potential utility advantage for OST settings, Playworks decided to use the *DESSA* as the behavior rating scale to pilot. The *DESSA* is a 72-item, strength-based behavior rating scale that assesses social and emotional competence of children (Smith et al., 2014). The *DESSA* yields an overall total score as well as scores across eight domains of social-emotional competence: Self-Awareness, Social Awareness, Self-Management, Relationship Skills, Goal-Directed Behavior, Personal Responsibility, Decision Making, and Optimistic Thinking. The *DESSA* takes 10-15 minutes to administer and can be completed by teachers, OST staff, parents, caregivers, and other important adults in the youth's life (LeBuffe, Ross, Fleming, & Naglieri, 2012). The *DESSA-Mini* (Naglieri, LeBuffe, & Shapiro, 2011/2014) is a brief (8-item) version of the *DESSA* that can be completed in 1 minute per child. The *DESSA* tools are psychometrically sound (Nickerson & Fishman, 2009; Naglieri, LeBuffe, & Shapiro, 2011), expert-reviewed (Tsang, Wong, & Lo, 2012; Merrell & Gueldner, 2010),

and practical (Denham, Ji, & Hamre, 2010; Haggerty, Elgin, & Woolley, 2011) for assessing the social-emotional competence of children (Maras, Thompson, Lewis, Thornburg, & Hawks, 2015).

The *DESSA* was piloted in two ways. First, a web-based administration of the *DESSA-Mini* was used at 27 Playworks sites. Second, a customized scannable paper administration of the full *DESSA* was used at 31 Playworks sites. Behavior rating scales were collected from OST staff, the child's primary teacher, and the child's primary caregiver in both pilots.

Selecting a Direct Observation Protocol

The selection of a direct observation tool of global character performance proved more difficult than the selection of a behavior rating scale. We did not identify a tool for a school-aged population through the *Handbook of SEL* (or any other means) that assessed individual children and was publically or commercially available with adequate user documentation to independently implement the assessment strategy. After several queries to various listservs asking for consultation in pursuit of this broad-band tool, it was discovered that work to fill this gap is underway at the Rush Neurobehavioral Center (RNBC). Rather than developing an assessment to directly observe the frequency of behavioral execution (or Social Emotional Competence, as it is termed in the *DESSA*), this team is working to develop a direct assessment of its theoretical precursor: social-emotional comprehension. Social-emotional comprehension is the ability to encode, interpret, and reason about social-emotional information (McKown, 2015).

The construct of social-emotional comprehension may be well aligned to many character education objectives, and may lead to specifically tailored, strength-based interventions for children receiving character education. Direct assessments are typically developed for indicated cases, and administered individually in the context of a clinical evaluation. Tools (like the one under development at RNBC) are needed to assess multiple dimensions of social-emotional comprehension, in groups, that can be scored and interpreted within practical program constraints (McKown 2015).

With the guidance and support of Dr. Nicole Russo-Ponsaran at RNBC, Playworks piloted a group administration of the Virtual Environment for Social Information Processing tool (*VESIP*; http://rnbc.org/2013/03/qa-vesip/) in an OST setting. *VESIP* is a computerized simulation in which children adopt the role of an avatar, interact with other avatars, respond to challenging social situations, and engage in social decision-making indicated by their real-time responses (Russo-Ponsaran, McKown, Johnson, Allen, & Knudsen, 2012). Although still under active development, Playworks conducted a feasibility test trying this highly innovative tool in an OST environment.

Results

Feasibility of using the *DESSA* as a Behavior Rating Scale in an OST setting

We informally observed three factors that affected feasibility of using the DESSA, namely administration format, regional director, and informant. On the one hand, the customized scannable paper forms were conducive to data collection in our OST environment. Paper surveys allowed for OST staff to complete assessments in action-oriented settings, for teachers to complete ratings "on the sideline" while OST staff were directing activities, and for parents to complete surveys sent home in backpacks. On the other hand, the web-based computerized assessments were conducive to formative assessment, as scores and reports were produced immediately, with the potential to shape practice in real-time. Technological developments may ultimately make this distinction obsolete if tablets become commonplace among OST program staff. For now, OST program leaders may want to select the administration format that reflects the needs that are most pressing in the OST setting.

The regional director also affected administration of the DESSA in two ways. First, the framing of the invitation by the regional director to participate mattered. Participants did not seem responsive to the request to complete the DESSA when regional directors used the term "pilot." Alternatively, when the term "evaluation" was used, participants seemed to be compliant with the request to complete forms. Second, when a regional director had a "can do" attitude, participants were likely to complete forms. Conversely, when the regional director or program staff approached the task with the attitude that teachers and parents are difficult to engage in OST programming, they were in fact, difficult to engage. It would be impossible to determine from this experience whether the attitude about engagement or the lack of actual engagement emerged first in specific communities; we can only hypothesize that assessment may be difficult without a mandate for participation or a plan that staff believe is achievable.

The feasibility of assessments differed by informant. OST staff were highly compliant (nearly 100%) with the request from their central office to assess students. When the local OST staff requested teacher participation, teacher completion rates varied from 38% - 72%, largely based upon aforementioned variables (e.g., the way the request was phrased, the administration format). Similarly, caregiver completion rates reached an impressive 70% in cases where there were strong program staff using customized scannable forms in an evaluation framework. In some regions, caregiver participation was limited since the *DESSA* is only available in English and Spanish.

Feasibility of using the *VESIP* for direct assessment in an OST setting

To use the *VESIP* within five Playworks sites, and assess approximately 15 youth per site, many resources were required. The test administrator needed to acquire mobile equipment which included a laptop, computer mouse, and headphones for each student; mobile hotspot devices and data plans that could sustain the number of computer connections needed; and power strips, extension cords, and suitable furniture to convert play spaces into adequate testing environments. Software needed to be pre-installed and tested on each computer, as did mobile connectivity to upload responses to a central server where scores would be collected and stored. All materials had to be transported to the OST sites and set up for testing in a room that was often used for another purpose only moments before the OST programming was to begin. Planned developments to *VESIP* involve evolution to a web-based platform such that no software downloads would be necessary and group administration through existing infrastructure (when available) would be possible. OST settings with access to computer labs (or even a small number of computers) could avoid many of these preparations and resource mobilization challenges.

There were clear advantages to conducting a group-administered direct assessment in an OST setting. The administration could be completed in a single OST session (for attending students), without relying on external informants. The OST staff were tremendous resources to the assessment process. The OST staff were trusted role models, having pre-existing relationships with students and the schools, who could readily problem solve environmental challenges, engage students in seated play during set-up, and role model and maintain warmth and attentiveness toward the guest administering the test. We imagine that it would ultimately be feasible to train select OST staff themselves to administer *VESIP* rather than rely on a highly credentialed mental health researcher. This could enable scale-up to more than five programs.

We have several other recommendations learned by adapting a direct assessment protocol to an OST setting. First, be prepared for parents to pick-up children early. It may be advisable to use 1 *VESIP* module (25 minutes) rather than 2 (45 minutes). Second, anticipate that students will desire to collaborate or compare their responses or time-to-completion with others during the assessment. Ask OST staff to strategically seat students together who are unlikely to talk to each other. This worked

better than simply spreading out the students as much as possible, since some OST spaces were actually too large to effectively monitor children's progress if all of the space was utilized. Third, remember that students have contextual expectations for OST environments that typically do not include testing. Particularly in the spring time, students may carry over test fatigue from the school day. The best results were achieved when following the advice of the administration guide and not overselling the avatar-enriched assessment as a "game", but also not calling it a "test", forging unnecessary associations for the children (the term "activity" was often used instead). Next, we recommend having a second quiet activity for students to complete in their seats, since the sequential start-time (each computer needs adult attention to start the administration) and the differential pacing of students led to different completion times. Once one student finished, the other students seemed more anxious to finish and join completed students in routine, active and engaging play. Finally, we recommend that test administrators leverage the unique skill sets of OST staff to make the administration successful. In this experience, the strengths of the OST staff is the factor that enabled and enhanced the use of a direct assessment in OST settings.

Discussion

Before we can answer key questions about the effectiveness of character education in OST environments, we need to enhance our capacity to measure the growth of youth character in the routine implementation of OST character education programs. Leming (2008) suggests that we approach this problem not exclusively as scientists in the quest for generalizable knowledge, but also as engineers, creating and testing high-quality solutions to respond to practical problems set in a social context.

Directors of OST programs, like Playworks, are increasingly being asked to determine and document the effectiveness of their programming. Despite demonstrated effects in other domains, and anecdotal reports from coaches, teachers, principals, and parents about the impact of Playworks on character development, empirical evidence of the effects of Playworks on youth character has so far been limited. Efforts to collect and evaluate this evidence have been constrained by the search for tools engineered for this specific purpose.

The field of Character Education provides some useful direction, naming criteria for the selection of assessment tools and maintaining a list of instruments that might meet these criteria. Recent acknowledgements of the convergence between the ethical and performance objectives of Character Education and the allied agenda of Social Emotional Learning allow for greater cross-disciplinary learning. The *Handbook of Social Emotional Learning* calls for the use of multi-modal assessment, including both direct observation approaches and behavior rating scales, when possible. Little guidance is provided, however, on the utility of various recommended approaches as applied to OST settings.

The authors of this paper could have piloted any number of strategies, but did not have the capacity to compete various methods with each other, given the routine demands placed on OST staff. Instead, the Director of Evaluation of Playworks, in consultation with researchers with expertise in assessment, selected two tools to test for feasibility within the organization during the 2014-2015 program year. There was no attempt by the authors to systematically collect feedback from stakeholders, but rather, to provide insights based on their experience of piloting these two tools. These insights include:

What features of social-emotional / character development assessment tools did we find attractive for use in our OST setting?

1. Tools that were strength-based and non-stigmatizing;
2. Tools with practical/limited requirements for assessor familiarity with the child;
3. Tools with consistent forms and scoring procedures across heterogeneous groups of children and adult informants;
4. Tools standardized and normed with OST program staff;
5. Tools that produce individual scores that can be used to tailor programming for individual children;
6. Tools that, on average, can be completed in about 2 minutes per child or on all children at the same time.

What lessons did we learn as we tried to use social-emotional / character development assessment tools in OST settings?

1. Paper behavior rating scales were easier to administer to adults than electronic behavior rating scales;
2. Electronic behavior rating scales were more useful than paper behavior rating scales for the purpose of formative assessments, given the expedience of report availability;
3. Response rates were highly dependent on the terminology used when behavior rating scales were distributed and the attitudes of regional staff toward the assessment process;
4. Behavior rating scales that rely exclusively on staff informants may be easier to collect in OST settings, but given the anticipated value of teacher and caregiver perspectives, modifications to assessment tools (e.g., translations into non-dominant languages) and data collection protocols (e.g., using paper forms) should be proactively considered;
5. Direct assessments, at this time, can be resource intensive to administer;
6. Direct assessments can engage students and be completed entirely during program time, although administrators should contingency plan for student absences and early departures, and thoughtfully consider how play spaces are modified into testing environments;
7. OST program staff can be a tremendous asset to overcoming test fatigue among students and helping the test experience reflect student's expectations of the OST setting.

The argument has been made that with an intense focus on effective youth development programs, we could reduce the incidence and prevalence of social, emotional, and behavioral health problems in the population by 20% within a decade (Hawkins et al., 2015), as well as promote the development of positive attributes in young people. Scaling up effective programs and practices could help millions of youth and save billions of dollars. Yet, we have not yet gone far enough to harness the power of such programs (Shapiro, 2015). Engineering tools for the rigorous and practical assessment of social-emotional and character development of children participating in youth development programs will help inform and evaluate our current practice, improve our programs, and ultimately develop children's character and capacities to thrive.

References

Afterschool Alliance. (2014). "America After 3PM." Retrieved from http://www.afterschoolalliance.org/documents/AA3PM-2014/AA3PM_National_Report.pdf.

Association for Supervision and Curriculum Development. (n.d.). Character education. *A Lexicon of Learning Online Dictionary*. Retrieved from http://www.ascd.org/Publications/Lexicon-of-Learning/C.aspx.

Berkowitz, M.W. (1998). *Primer for Evaluating a Character Education Initiative*. Washington, DC: Character Education Partnership.

Berkowitz, M.W., Battistich, V.A., & Bier, M.C. (2008). What works in Character Education: What is known and what needs to be known. In L. Nucci, & D. Narvaez (Eds.), *Handbook of Moral and Character Education*. New York: Routledge.

Berkowitz, M.W., & Bier, M.C. (2005). *What Works in Character Education: A Research-driven Guide for Educators*. Washington, DC: Character Education Partnership.

Catalano, R.F., Toumbourou, J.W., & Hawkins, J.D. (2014). Positive youth development in the United States: History, efficacy, and links to moral and character education. In L. Nucci, D. Narvaez, & T. Krettenauer (Eds.), *Handbook of moral and character education (2nd ed.).* (pp423-440). New York and London: Routledge.

Krettenaurer. (Eds.), *Handbook of Moral and Character Education (2nd ed)*. New York: Routledge. Character Education Partnership. (2008). *Performance Values: Why They Matter and What Schools Can Do to Foster Their Development*. Washington, DC.

Denham, S.A. (2015). Assessment of SEL in Educational Contexts. In J.A. Durlak, C.E. Domitrovich, R.P. Weissberg, & T.P. Gullotta (Eds.), *Handbook of Social and Emotional Learning: Research and Practice*. New York: Guilford.

Denham, S.A., Ji, P., & Hamre, B. (2010). *Compendium of preschool through elementary social-emotional learning and associated assessment measures*. Chicago, IL: Collaborative for Academic, Social, and Emotional Learning.

Doll, E., & Elliott, S.N. (1994). Consistency of observations of preschoolers' social behavior. *Journal of Early Intervention, 18*(2), 227-238.

Elias, M.J., Kranzler, A. Parker, S.J., Kash, V.M., & Weiessberg, R.P. (2014). The complementary perspectives of social and emotional learning, moral education, and character education. In L. Nucci, D. Narvaez, & T. Krettenaurer (Eds.), *Handbook of Moral and Character Education (2nd ed)*. New York: Routledge.

Elliott, S.N., Frey, J.R., & Davies, M. (2015). Systems for assessing and improving students' social skills to achieve academic competence. In J.A. Durlak, C.E. Domitrovich, R.P. Weissberg, & T.P. Gullotta (Eds.), *Handbook of Social and Emotional Learning: Research and Practice*. New York: Guilford.

Epstein, M.H. (2004). *Behavioral and emotional rating scale (2nd ed.).* Austin, TX: PRO-ED.

Fortson, J., James-Burdumy, S., Bleeker, M., Beyler, N., London, R.A., Westrich, L., Stokes-Guinan, K., & Castrechini, S. (2013). Impact and implementation findings from an experimental evaluation of Playworks: Effects on school climate, academic learning, student social skills, and behavior. Retrieved from: http://www.rwjf.org/content/dam/farm/reports/evaluations/2013/rwjf405971.

Gresham, F.M., & Elliott, S.N. (2008). *Social Skills Improvement System - Rating Scales*. Minneapolis, MN: Pearson Assessments.

Gullotta, T.P. (2015). After-school programming and SEL. In J.A. Durlak, C.E. Domitrovich, R.P. Weissberg, & T.P. Gullotta (Eds.), *Handbook of Social and Emotional Learning: Research and Practice*. New York: Guilford.

Haggerty, K., Elgin, J., & Woolley, A. (2011). Social-emotional learning and school climate assessment measures for middle school youth. Retrieved from the Raikes Foundation website: http://www.raikesfoundation.org/

Hawkins, J.D., Jenson, J.M., Catalano, R.F., Fraser, M.W., Botvin, G.J., Shapiro, V.B., Bender, K.A., Brown, H., Beardslee, W., Brent, D., Leslie, L.K., Rotheram-Borus, M.J., Shea, P., Shih, A., Anthony, E.K., Haggerty, K.P., Gorman-Smith, D., Casey, E., Stone, S., & the Coalition for Behavioral Health. (2015). Unleashing the Power of Prevention. *American Academy of Social Work & Social Welfare Grand Challenge Initiative*. Paper No. 10.

LeBuffe, P.A., Ross, K.M., Fleming, J.L., & Naglieri, J.A. (2012). The Devereux Suite: Assessing and promoting resilience in children ages 1 month to 14 years. In S. Prince-Embury & D. Saklofske (Eds.). Translating Resiliency Theory for Application with Children, Youth, and Adults.

LeBuffe, P.A., Shapiro, V.B., & Naglieri, J.A. (2009/2014). *The Devereux Student Strengths Assessment* (DESSA) *Assessment, Technical Manual, and User's Guide*. Charlotte, NC: Apperson, Inc.

Leming, J.S. (1993). In search of effective character education. *Education Leadership, 51(3)*, 63-71.

Leming, J.S. (2008). Research and practice in moral and character education: Loosely coupled phenomena. In L. Nucci, & D. Narvaez (Eds.), *Handbook of Moral and Character Education*. New York: Routledge.

London, R.A., Westrich, L., Stokes-Guinan, K., & McLaughlin, M. (2015). Playing fair: The contribution of high-functioning recess to overall school climate in low-income elementary schools. *Journal of School Health, 85*(1): 53–60.

Maras, M.A., Thompson A.M., Lewis, C., Thornburg, K. & Hawks, J. (2015). Developing a tiered response model for social-emotional learning through interdisciplinary collaboration. *Journal of Educational and Psychological Consultation, 25,* 1-26.

McKown, C. (2015). Challenges and opportunities in the direct assessment of Children's Social and Emotional Comprehension. In J.A. Durlak, C.E. Domitrovich, R.P. Weissberg, & T.P. Gullotta (Eds.), *Handbook of Social and Emotional Learning: Research and Practice*. New York: Guilford.

Merrell, K.W. (1993). Using Behavior Rating Scales to Assess Social Skills and Antisocial Behavior in School Settings: Development of the School Social Behavior Scales. *School Psychology Review, 22*(1) 115-133.

Merrell, K.W. (2011). *Social and emotional assets and resilience scales (SEARS)*. Lutz, FL: Psychological Assessment Resources.

Merrell, K.W., & Gueldner, B.A. (2010). *Social and Emotional Learning in the Classroom: Promoting Mental Health and Academic Success*. New York: Guilford Press.

Naglieri, J.A., LeBuffe, P.A., & Shapiro, V.B. (2011/2014). *The Devereux Student Strengths Assessment - Mini (DESSA-Mini) Assessment, Technical Manual, and User's Guide.* Lewisville, NC: Kaplan.

Naglieri, J.A., LeBuffe, P.A., & Shapiro, V.B. (2011). Universal screening for social emotional competencies: A study of the reliability and validity of the DESSA-mini. *Psychology in the Schools, 48*(7), 660-671.

Naglieri, J.A., LeBuffe, P.A., & Shapiro, V.B. (2013). Assessment of social-emotional competencies related to resilience. In S. Goldstein & R. Brooks (Eds.), Handbook of Resilience in Children. NY, NY: Kluwer/Academic Press.

National Clearinghouse on Families & Youth. (2007) Putting positive youth development into practice: A resource guide. Silver Spring, MD: U.S. Department of Health and Human Services Administration for Children and Families, Administration on Children, Youth, and Families, & Family and Youth Services Bureau.

National Afterschool Association. (2011). Core Knowledge and Competencies for Afterschool and Youth Development Professionals. Retrieved from: http://naaweb.org/images/PDFs/NAA_CKC_Blue_Cover.pdf.

Nickerson, A.B., & Fishman, C. (2009). Convergent and divergent validity of the Devereux Student Strengths Assessment. *School Psychology Quarterly. 24*(1), 48-59.

Nucci, L., Narvaez, D., & Krettenauer, T. (2014). *Handbook of Moral and Character Education (2nd ed)*. New York: Routledge.

Russo-Ponsaran, N.M., McKown, C., Johnson, J.K., Allen, A., & Knudsen, K. (2012). Usability & Likability of the Virtual Environment for Social Information Processing (VESIP) for Children with and without Autism Spectrum Disorders. Presented at the International Society for Autism Research: Toronto: Ontario.

Sipos, R. & Maupin, L (2010). 11 Principles of Effective Character Education. Character.org: Washington, DC.

Smith, G.T., Shapiro, V.B., Sperry, R.W., & LeBuffe, P.A. (2014). A strengths-based approach to supervised visitation in child welfare. *Child Care in Practice, 20*(1), 98–119.

Shapiro, V.B. (2015). Resilience: Have we not gone far enough? A response to Larry Davis. *Social Work Research, 39*(1): 7-10.

Tsang, K.L.V., Wong, P.Y.H & Lo, S.K. (2012). Assessing psychosocial well-being of adolescents: A systematic review of measurement instruments. *Child: Care, Health and Development. 38*(35), 629-646.

Weissberg, R.P., Durlak, J.A., Domitrovich, C.E., & Gullotta, T.P. (2015). Social and Emotional Learning: Past, present, and future. In J.A. Durlak, C.E. Domitrovich, R.P. Weissberg, & T.P. Gullotta (Eds.), *Handbook of Social and Emotional Learning: Research and Practice*. New York: Guilford.

ISSN 2325-4009 (Print); ISSN 2325-4017 (Online)