QTURN WHITE PAPER #3

Realist(ic) Evaluation Tools for OST Programs

The Quality-Outcomes Design and Methods (Q-ODM) Toolbox

Updated 07/13/21

"By making the study of facts subservient in advance to its final purpose of a mathematical play with symbols, not only does it fail to stimulate progress in the analysis of these facts, but actually obstructs it" (Znaniecki, 1934, p. 231; from Chirkov & Anderson, 2018a, p. 725).

Contents

Citation: Peck, S. C., & Smith, C., (2020). *Realist(ic) Evaluation Tools for OST Programs: The Quality-Outcomes Design and Methods (Q-ODM) Toolbox.* [White Paper #3]. QTurn.

‘

This paper is part of a series: White Paper 1 – *Socio-Emotional Skills, Quality, and Equity* (Peck & Smith, 2020b) – provides a translational framework for understanding the key parts of an SEL skill set. White Paper 2 – *Measuring Socio-Emotional Skill, Impact, and Equity Outcomes* (Smith & Peck, 2020a) – provides guidance for selecting feasible and valid SEL skill measures. White Paper 3 – *Realist(ic) Evaluation Tools for OST Programs* – integrates the SEL framework and measures with a pattern-centered approach to both CQI and impact evaluation. White Paper 4 – *Citizen Science and Advocacy in OST* (Smith & Peck, 2020b) – presents an alternative evidence-based approach to improving both the impact and equity of OST investments.

Summary

Socio-emotional learning (SEL) skills are a partial but necessary cause of children's developmental outcomes, and SEL skill growth is a key objective for nearly all out-of-school time (OST) programs. The *Quality-Outcomes Design and Methods* (Q-ODM) toolbox holds an integrated set of tools to measure and model children's SEL skills, including how they change during, and in response to, OST programs (e.g., afterschool, school-age child care, workforce and career preparation, arts, sports). The Q-ODM toolbox helps organizational managers and evaluators to feasibly and cost-effectively adopt *pattern-centered* measures and models that produce actionable information for both continuous quality improvement (CQI) and impact evaluation.

The Q-ODM toolbox addresses practical questions about SEL skills and skill growth, such as: What is high-quality SEL support? How much SEL skill change does our program cause in each cycle? How much program quality does it take for stressed children to fully engage? Does our work create equity effects? The tools are divided into three groups: Design Tools, Analytic Tools, and Feedback Tools. These tools increase dramatically the value of CQI feedback for staff and the power of the analytic models used to evaluate program impact and equity effects for participating children. The Q-ODM toolbox was designed to empower internal and local evaluators to conduct rigorous and meaningful impact evaluations using existing resources (e.g., while they are implementing their current CQI systems). These tools will be particularly welcomed by evaluators currently struggling with *positivist* thinking and methods. [1]

I. Introduction

The *Quality-Outcomes Design and Methods (Q-ODM) toolbox* is a suite of tools for measuring and modeling out-of-school time (OST) program quality, SEL skill growth, and SEL skill transfer outcomes for purposes of both continuous quality improvement (CQI) and impact evaluation. The Q-ODM toolbox helps program managers and evaluators (a) communicate with stakeholders about design, measurement, and analysis issues; (b) implement Q-ODM tools using internal resources; and (c) inform staff, management, and funder decision-making. The results of applying Q-ODM tools to the kinds of data typically generated by CQI and evaluation projects generally support the conclusion that: For all OST settings and systems (e.g., afterschool programs, school age child care, and workforce and career preparation), the quality of SEL supports and children's [2] SEL skills are partial but necessary causes of a wide array of child outcomes (e.g., 3rd grade reading, 8th grade algebra, prosocial behavior, and positive mental health). Given that Q-ODM tools are ideally suited for identifying and modeling diversity (e.g., detecting equity effects), we typically find that the impacts of high-quality SEL supports are particularly evident for children who enter OST programs with relatively under-developed SEL skill sets.

We focus on SEL equity because children who cannot fully engage during program sessions – due to self-regulation challenges, the setting's low quality, or both – are unlikely to learn at the same rate as their better self-regulated and better-supported peers. Focusing on diversity (e.g., taking account of children with the full range of SEL skill sets, as opposed to focusing models on children with modal or average SEL skill sets) provides clear information about equity effects. Clear information about equity effects, in turn, provides practitioners with the kind of detailed information about children's SEL skill sets that is necessary for adjusting their practices to meet the needs of all children whom they serve. In particular, given that SEL skills have a compounding effect on many outcomes – that is, there is a *dynamic complementarity* (Heckman, 2007) between SEL and other skills, such that SEL skills beget other types of skill – it is critical for practitioners to understand every child's SEL skill set. The fundamental equity question is: Does the setting support all children to successfully engage with and learn from program activities? Recent work on trauma-informed practices and mindfulness can also be reframed in terms of SEL equity, and the Q-ODM toolbox provides methods for incorporating and applying information and evidence from these areas of clinical practice.

The fundamental equity question is: Does the setting support all children to successfully engage with and learn from program activities? Recent work on trauma-informed practices and mindfulness can also be reframed in terms of SEL equity, and the Q-ODM toolbox provides methods for incorporating and applying information and evidence from these areas of clinical practice.

Because instructional quality and SEL skills are so important to achieving the goals of OST programs, it is unfortunate that the most widely used approaches to measurement and evaluation – *positivist* theory and methods (described below) – are poorly suited to measuring and modeling quality and SEL skill. Solutions are out there (e.g., in the form of alternative methods and local expressions of program quality that actually help children grow their SEL skills), but because positivist approaches to OST measurement and evaluation generally ignore them, the field rarely expands its toolbox, learns from its own exemplars, or attracts the funding necessary to document and promote high performance.

The Positivist Challenge

We use the term "positivist" where referring to a pervasive knot of assumptions and methods (e.g., behaviorist theory, psychometrics, linear statistical models, and counterfactualism)[3] that represent a mainstream social science consensus about valid measurement and evaluation. Although there are many long-established facets to the critique of this consensus, two big ones are: (a) the misfit between analytic models and reality, or *construct invalidity,* and (b) the neglect of children's full range of mental skills, or *consequential invalidity*.

If SEL behavioral skills are a combined function of schemas, beliefs, and awareness (Smith & Peck, 2020a), but measures and models of SEL mental skills include reference to only beliefs, then we are probably neglecting about two-thirds of what is important for understanding children's subjective experience of program quality, SEL behavioral skills, and positive youth development more generally.

The *misfit between analytic models and reality* refers to our observation that theories and measures of OST program quality and children's SEL skills often fail to focus on relevant causal forces within OST programs or children. In particular, they typically fail to distinguish clearly between children's SEL mental and behavioral skills and among different aspects of mental skill. As a result, they rarely represent sufficiently either the various parts of SEL skill or children's integrated set of SEL skills as a whole. For example, in our view, integrated SEL skill sets almost always include three key aspects of SEL mental skills (i.e., schemas, beliefs, & awareness) that combine to produce SEL behavioral skill (Smith & Peck, 2020a). Further, even if the measurement items used to assess SEL skills are valid, statistical models using these data typically fail to adequately represent children's SEL skill sets because the quantities they model tend to be aggregate, sample-level abstractions (e.g., means, partial correlation coefficients, effect size estimates) that accurately represent few, if any, of the children whose responses were used to create those statistical abstractions (Cairns, 1986; Lamiell, 2013; Richters, 1997; Rose et al., 2012; Uher, 2019).

The implication of construct invalidity for the OST field is that the information produced by positivist-oriented measures and models is not very useful for either the CQI efforts of program staff and mangers or the evaluative decisions that managers and funders need to make as a part of good leadership. In other words, just as you should not stop looking for your keys after searching under the light, you should not give up on finding useful program impact information after obtaining nil effects from positivist methods. Instead, keep in mind that despite whatever granular indicator-level information about settings and children you may have, neither the measurement models by which they are combined into scales nor the linear statistical models used to describe their interactions necessarily represent clearly the realities of program settings, children's skills, or their interactions. This is clearly problematic, but it is probably not the worst problem.

The *neglect of children's full range of mental skills* (or, in too many cases, the complete neglect of any and all mental skills) is stubbornly persistent and reflects some of the decades-old criticisms of positivist assumptions and methodology (Danziger, 1990; Mackenzie, 1977). Such neglect is one of the profound consequences of the misfit between (a) analytic models of the relations between program quality and children's SEL skills and (b) the presumed, or theoretically specified, underlying reality of program quality and children's SEL skills. In short, if SEL behavioral skills are a combined function of

schemas, beliefs, and awareness (Smith & Peck, 2020a), but measures and models of SEL mental skills include reference to only beliefs, then we are probably neglecting about two-thirds of what is important for understanding children's subjective experience of program quality, SEL behavioral skills, and positive youth development more generally.

This kind of benign neglect of children's mental skills and, particularly, the full range of mental skills, is an abiding problem for an OST field anchored and animated by the developmental concept that children's mental skills, enacted in context, are the primary drivers of growth and change, which is the causal theory at the core of the positive youth development approach (Eccles & Gootman, 2002; Larson, 2000). The mismeasure and mismodeling of the full range of children's SEL skills means that the actual parts of SEL skill that practitioners must understand to be fully responsive to each child's SEL history – and to achieve more equitable outcomes – remains invisible and neglected. As a result, despite our best intentions, the representation, understanding, and appreciation of children's mental skills continues to be impoverished.

Given the default positivist-type mode of inquiry that dominates the OST field, the neglect of the full range of children's mental skills plays itself out insidiously in real OST settings. For example, where the complete, holistic structure of children's SEL mental skills are not reflected in the corresponding data and analytic models, valid decisions about setting quality, children's SEL skills, and their interactions remain unlikely. In particular, the pervasive neglect of children's attachment schemas (and corresponding emotions) and executive attention abilities is a major challenge to the rational operation and evaluation of OST programs. Elsewhere (Smith & Peck, 2002b), where discussing Dynarski et al.'s (2001) evaluation of 21st Century Community Learning Center programs, we provide a more detailed discussion of positivist "methodolatry" (Daly, 1973; Pepper, 1942) and the neglect of the full range of children's mental skills. In short, evidence of the systemic disregard of how teachers and students feel about the conditions of learning is reflected in the ample literature documenting and condemning the failed "No Child Left Behind" accountability policy (Cohen, 2006; Ho, 2008; Husband & Hunt, 2015; Ladd, 2017).

The Q-ODM tools reflect the assumption that children have SEL mental skills that are causes of their behavior in OST and other settings.

Realist(ic) Tools

Alternative pathways out of the stranglehold of positivist assumptions and methods vary widely and lack consensus, due largely to a complex web of "conceptual fallacies" (Uher, 2020, p. 1), but we are making progress. The Q-ODM toolbox provides a *realist*[4] and *pattern-centered*[5] alternative for applied measurement and evaluation that joins a growing number of alternatives to social science business as usual (e.g., Arocha, 2020; Burton-Jones & Lee, 2017; McGill et al., 2021; Uher, 2019, 2021). Numerous examples of CQI and impact evaluation using the Q-ODM approach are available on the QTurn website (www.qturngroup.com; e.g., Lindeman et al., 2019; Peck & Smith, 2020a; Smith & Peck, 2019). These tools are anchored generally in philosophical realism, a set of assumptions about what's real and what can be known about such "real" things that contrast substantially with positivist assumptions.

Of central importance to the OST field are the ideas that individual children have mental skills, they use these skills to actively participate in their own learning and development, and the full scope of these skills – the holistic pattern – must be adequately reflected in our theories, measures, and models in order to accurately and ethically describe individual change. The Q-ODM tools reflect the assumption that children have SEL mental skills that are causes of their behavior in OST and other settings. These

mental skills are conceived of as integrated systems composed of several different aspects of mental functioning (i.e., schemas, beliefs, & awareness) that exist within every biologically-intact person, enable behavioral skills, and can be assessed, more or less accurately, using properly-aligned SEL measures.

When the parts and patterns of SEL skill are reflected in theory and measures, the accuracy and meaningfulness of data about program quality and SEL skill - *and all subsequent manipulations and uses of the data* – are dramatically improved. The Q-ODM tools reflect a *methodological realism*[6] that sets the rules for these subsequent manipulations. These tools focus on procedures for selecting or composing theoretically-specified measurement items (or, indicators), measuring SEL skills and skill growth, and modeling program impact and equity effects. These pattern-centered tools are anchored in the assumptions of psychological realism[7] and meet the requirements of methodological realism.

The Q-ODM tools are also "realistic" in the sense that they are feasible and cost effective. In a previous paper (Smith et al., 2019b), we suggested that afterschool systems use the Q-ODM toolbox to "measure once, cut twice" (p. 3); that is, simultaneously set up basic internal data systems for *both* continuous quality improvement (CQI) *and* impact evaluation. Using data for CQI purposes is generally a noncontroversial, evidence-based practice in both OST and other fields (Lester, 2018; Smith et al., 2012). However, many people believe that the only valid way to estimate the "impact" of OST program quality on child outcomes is by using a design with random assignment of individuals or programs to experimental and comparison groups. We do not restrict our definition of "impact" to experimental designs or other counterfactualist approaches.[8] Instead, we provide an alternative and rigorous pathway for high-capacity organizations to produce and replicate their own impact studies, with the samples at hand.

We do not restrict our definition of "impact" to experimental designs or other counterfactualist approaches.

II. Q-ODM Toolbox

The Q-ODM toolbox consists of nine tools in three categories: *Design Tools* include Theory of Change, Object Alignment, and Question Logic. *Analytic Tools* include Profile Measures, Growth Indicators, and Impact/Equity Models. *Feedback Tools* including Behavior Analogy, Visual Holism, and Benchmarks. The Tools were designed as an integrated set of methods, but each tool can also serve an independent purpose in the process of using evidence to improve children's outcomes.

children exposed to high-quality instructional practices at the point of service are more likely than children exposed to lower-quality practices to undergo socio-emotional skill growth.

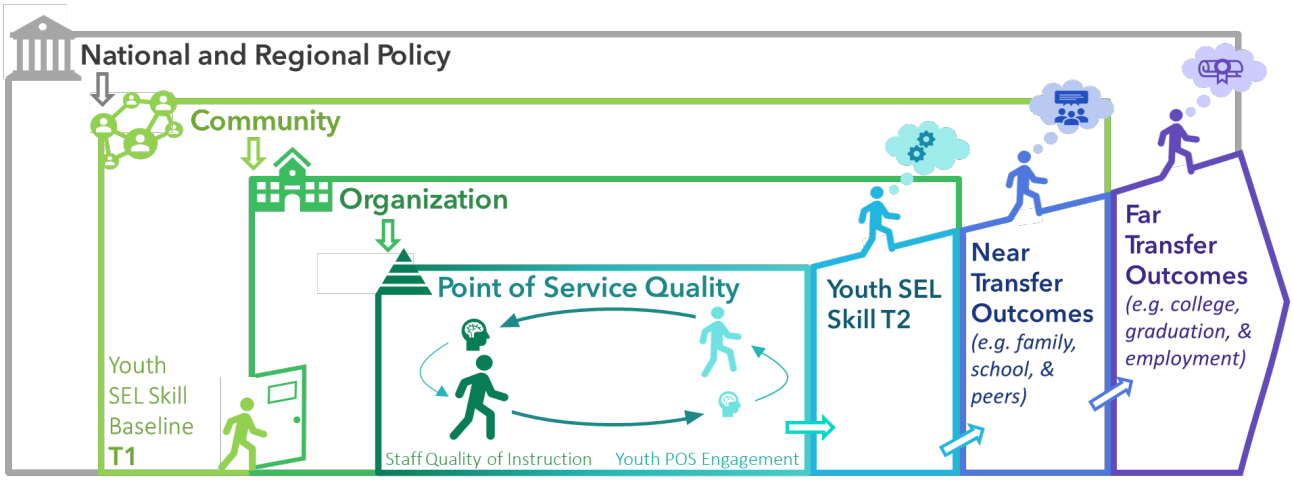
Design Tools

The Theory of Change makes it easier to have internal and external conversations about (a) how (and how much) individual skill growth occurs in OST settings and (b) how program quality and SEL skill growth are causes of subsequent outcomes. The remaining design tools (i.e., Object Alignment and Question Logic) support internal conversations about how to measure and model SEL skill change; for example, how to align measures to the theory of change to generate answers to specific questions about antecedent causes, individual change, and subsequent transfer outcomes. The Design Tools are focused on bringing stakeholders along as consumers and producers of information that flows from using the tools.

Theory of Change

In a previous paper (Smith et al., 2019a), we introduced the Multilevel Person-in-Context~*neuroperson* (MPC*n*) theory of change summarized in Figure 1 (see, also, Peck & Smith, 2020b). Given the many person-in-context models found in the psychological literature, we added the term *neuroperson* to both emphasize mental skills and convey our view that the three fundamental aspects of SEL mental skill (i.e., schemas, beliefs, and awareness) are centered in three different areas of the brain (i.e., subcortical, cortical, and prefrontal) (cf. Peck, 2007, 2009; Roeser et al., 2006; Roeser & Peck, 2009). In this view, and as shown in Figure 1, children are multilevel systems who are embedded in multilevel social systems. More specifically, children exposed to high-quality instructional practices at the point of service are more likely than children exposed to lower-quality practices to undergo socio-

Figure 1. Multilevel Person-in-Context~neuroperson (MPCn) Theory of Change.



emotional skill growth. These bourgeoning skills can then be applied in other settings, causing a wide variety of positive near- and far-transfer outcomes (e.g., 3rd grade reading, post-secondary entry). This cascade of causes and effects of children's SEL skill growth flows through OST program quality and children's mental engagement at the point of service.

The reciprocal dynamics between staff and children at the point of service are key drivers of the SEL skill growth process. Children who enter OST settings with lower SEL skill and receive lower-quality instruction are unlikely to generate and sustain mental engagement during program activities, putting them at further risk of SEL skill stagnation or loss. In contrast, children who enter OST settings with lower SEL skill and receive high-quality instruction are likely to generate and sustain mental engagement during program activities; and, as a result, they are much more likely to experience SEL skill growth and transfer those skills to both contemporaneous (i.e., near transfer) and future (i.e., far transfer) settings. The Q-ODM toolbox can be used to evaluate program performance in terms of these basic assumptions.

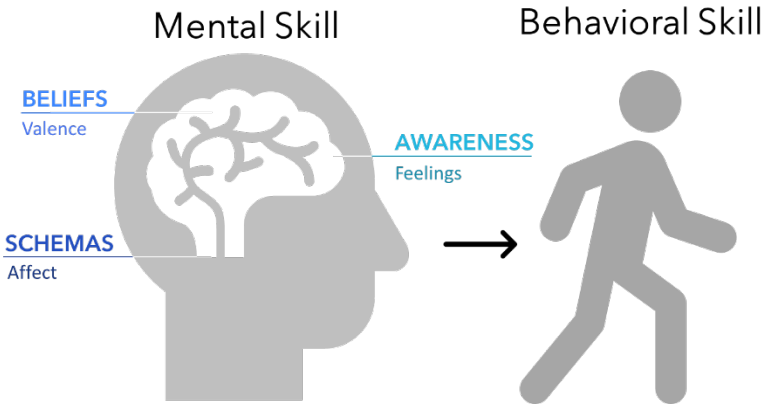
Producing accurate, informative, and useful data requires measures that are aligned to the real causes and effects in persons and settings. Analytic models that use those measures to describe the relations among features of persons and settings are meaningful only to the extent that measures adequately represent the intended objects.

Figure 2 shows the three parts of an integrated SEL skill set (i.e., schemas, beliefs, and awareness) that cause child behavior in any setting. Note that *emotion* processes are influenced separately by each part of an integrated SEL skill set – as indicated by the terms *affect*, *valence*, and *feelings* – which means that every SEL skill set has emotional implications. Additional details about the MPC*n* theory of change and corresponding Neuroperson model can be found in White Papers 1 (Peck & Smith, 2020b) and 2 (Smith & Peck 2020a).

Object Alignment

Producing accurate, informative, and useful data requires measures that are aligned to the real causes and effects in persons and settings. Analytic models that use those measures to describe the relations among features of persons and settings are meaningful only to the extent that measures adequately represent the intended objects. The following paragraphs describe some of the important 'objects' that exist at different levels of the typical OST program system and provide examples of how measures can be aligned to those objects. Similar measure-object alignment examples can be found in previous work (e.g., Smith, McGovern, Peck, et al., 2016, Appendix D; Smith et al., 2017).

Figure 2. The Neuroperson Model.



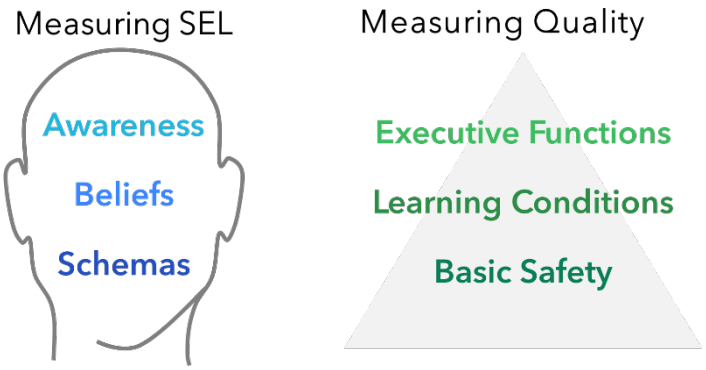
Organization. At the organization level, the objects of interest are management practices. Do managers implement the CQI process and shape a collaborative

organizational culture that is comfortable with transparency about instruction? The purposes of performance measurement cannot be served if the CQI cycle is not implemented, and the CQI cycle will not be very effective if staff do not feel safe sharing with each other. We have used the following measures to align with management practices: (a) *CQI implementation fidelity*, (b) *vertical communication* between managers and staff, and (c) *horizontal communication* between staff. We recommend using staff reports, as opposed to manager reports, as the source of the data about management practices [9] (Smith, McGovern, Peck, et al., 2016).

Point of Service. Staff instructional practices and youth engagement [10] are the objects of interest at the point of service and the primary anchors for objectivity in a set of CQI performance measures. Figure 3 shows alignment between the three parts of SEL skill and three widely used domains of program quality: *Basic Safety, Conditions of Learning, Executive Engagement*. There are numerous instruments available [11] for assessing the quality of instructional practices at the point of service, most of which include dimensional scores (or reconfigurable items) aligned to practices that target the parts of SEL mental skill.

Intra-Person. According to the neuroperson model (see Figure 1), the main parts of a person's SEL skill set exist at different levels of the self-system, such that the objects of measurement must be considered in different terms. For example, mental objects at different levels of the self-system (e.g., schemas, beliefs, and awareness) change in different ways and rates, even where they are the parts of an integrated skill that functions holistically as a uniquely structured higher-order 'object' within a given child. Aligning measurement items or scales to SEL skills requires distinguishing clearly among the various parts (schemas, beliefs, awareness, behavior) and types (functional, optional) of SEL skill at two or more timepoints. These issues are the specific focus of White Paper 2 (Smith & Peck, 2020a).

Figure 3. Parts of SEL Mental Skill.



Measures of academic achievement, subject-matter performance (grades), and other school-related behavior (e.g., expulsions, suspensions) are potential sources of evidence of successful SEL skill transfer from OST to school day settings. Aligning measures to two or more objects in the theory of change creates the opportunity to develop benchmarks for performance that are related to outcomes (e.g., the benchmarked levels of instructional quality and SEL skill growth at which transfer outcomes occur). Unfortunately, conventional (e.g., positivist) evaluation approaches rarely include measures that are aligned well to the key objects within the point of service or the developing child; as a result, they rarely provide meaningful information about the impact of instructional quality on SEL skill growth or the impact of SEL skill growth on transfer outcomes.

Question Logic

Q-ODM tools address practical questions about 'how' and 'how much' by using measures aligned to the theory of change. The logic of the causal cascades represented in the theory is that higher-quality programming, instructional practices, and children's mental engagement cause growth in children's integrated SEL skill sets that influence transfer outcomes. [12] The evaluation questions in Table 1 reflect

this guiding logic and its emphasis on instructional quality and SEL skill growth during the OST program period. Answers to the first three questions yield valid CQI feedback known to improve quality in OST programs (Smith et al., 2012). Answers to the second three questions yield estimates of program impact and equity effects. [13]

It is important to see the logic of the causal cascade in the questions, because the real objects of interest (e.g., the specific practices and skills and the local terms used to identify these objects) tend to differ across organizations and networks. However, by following the logic, users can fill in their own specific practices and terms for quality, skill growth, transfer outcomes, and equity.

.

Table 1. CQI and Impact Evaluation Question Logic.

CQI	(1) What is the prevalence of low-/high-quality within and across settings? (2) What is the prevalence of children with low-/high-skill at baseline within and across settings? (3) What is the prevalence of children following SEL skill growth, stability, or decline pathways, from baseline to follow-up, within and across settings?
Impact Evaluation	(4) Does exposure to higher-quality programs cause more skill change than exposure to lower-quality programs? (5) Do children with lower SEL skills at baseline gain as much or more than children with higher SEL skills at baseline where exposed to high quality? (6) Do children who were exposed to high quality, or who experienced SEL skill growth, demonstrate improved transfer outcomes (e.g., better grade, less recidivism)?

Analytic Tools

The Q-ODM analytics tools help evaluators and technical staff improve the validity of SEL skill measures, create SEL skill growth indicators from multi-timepoint data, and model impact and equity effects. Pattern-centered methods are ideally suited for evaluating OST programs because they are anchored in strong theory (i.e., MPCn) and provide flexibility to integrate information from diverse measures, timepoints, and levels/units of analysis. They also allow for tests of the relations between program features that are less likely to be "under powered" or uninterpretable with small samples, a constraint often confronting OST evaluations based in positivist theory and methods.

Profile Measures

In pattern-centered measurement models (Magnusson, 2003), a person's score on a given measure (e.g, variable, scale) gets its meaning not by reference to other people's scores on that same measure (as in conventional psychometrics) but by reference to their own scores on other theoretically relevant measures. Assuming measures that have been properly aligned to objects at a given level, the pattern of scores across a set of theoretically-relevant measures constitute a holistic *measurement model* for those integrated objects, typically setting quality or SEL skills. In practice, to create profiles, indicators need to be aligned (composed, scaled) with the parts of SEL skill and then integrated in a profile measure.

Pattern-centered methods are ideally suited for evaluating OST programs because they are anchored in strong theory and provide flexibility to integrate information from diverse measures, timepoints, and levels/units of analysis.

Align Indicators with Objects. Often, SEL skill measures must be composed from the items at hand [14] so that the parts of mental skill, behavior, and setting quality can be clearly distinguished. In general, greater "granularity" of SEL indicators improves the explanatory power of data and analytic model (as well as the precision of measures, as in interrater reliability [15]). Measurement item *response scales* (e.g., 1 = never, 3 = sometimes, 5 = always) can also be more accurately aligned with specific objects, which can in many cases increase dramatically the usefulness of the data for both CQI and impact evaluation.

Like reconfiguring items and scales to better differentiate the parts of SEL skill, the process of *quasi-absolute scaling* can also increase the accuracy and interpretability of data [16] by engaging local expertise to define locally-meaningful levels of the item response scale. For example, the minimum level of SEL skill necessary for a child to successfully achieve program goals is likely different for emotion management (most of the time) compared to planfulness (some of the time). [17]

Similarly, the frequency at which an instructional practice is likely to produce tangible and visible changes in children's SEL skills may also differ by practice. The critical difference between a practice that achieves required levels of fidelity – suggesting that positive effects should follow – and various forms of non-implementation that are unlikely to produce any effect can (absent empirical testing) only be known for each indicator by asking a local expert. The utility value of the data depend decisively on the meaning of the response-scale values, as does the clarity and interpretability of visual representations of the data.

Integrate Indicators in a Profile. Valid indicators for the parts of SEL skill can be integrated to create profile measures of the integrated skill set that each child brings to the setting. The procedures for constructing pattern-centered measurement models (i.e., creating level and time-specific profiles) include: (a) composing valid indicators, (b) imputing missing data values, (c) temporarily removing multivariate outliers, (d) identifying an optimal set of profiles, (e) reassigning individual cases (including multivariate outliers) to the profile group that best matches their particular profile, and (f) assessing the reliability and validity of the resulting profile variable (Bergman et al., 2003; Vargha et al., 2015). These procedures are described elsewhere in more detail (e.g., Peck & Smith, 2020a, App. A). Point-in-time profiles constructed from valid SEL skill indicators constitute pattern-centered measurements models of integrated SEL skill sets and provide the necessary basis for accurately modeling SEL skill growth pathways. [18]

Growth Indicators

Our pattern-centered approach to measuring SEL skill *growth* differs substantially from mainstream approaches [19] and addresses a number of pressing challenges to the validity of positivist measurement and skill growth models. We tend to think about SEL skill change in qualitative, or nonlinear, terms rather than quantitative, or linear terms, but both perspectives are generally brought to bear at different points in the process. For example, we need to determine the extent to which an individual child's multivariate profile pattern at one point in time matches their profile at other points in time, and quantitative estimates can be useful during this process. [20] We also have to contextualize individual-level change by reference to the sample-level changes in profile patterns so that, for example, the "high-skill" profile at T1 has the same meaning as the high-skill profile at T2. During this process, we also identify *vanishing* profiles (e.g., a "very low" profile at T1 that is found at no later time) and *emerging* profiles (e.g., an "exemplary" profile that appears at a later time that did not exist at T1). In other words, individual-level changes (e.g., children's SEL growth pathways) are defined in relation to sample-level profile patterns that have been aligned across timepoints to produce an 'objective framework' for understanding individual-level pathways of stability and change.

We tend to think about SEL skill change in qualitative, or nonlinear, terms rather than quantitative, or linear terms…

With the sample-level profile structures aligned across timepoints, an individual-level pathway indicator is created for each child indicating the type of change in SEL skills that occurred for each child. Although this procedure generally reveals many different pathways, we typically focus on three simple pathway patterns: Growth, Stability, and Decline. For example, if a child was in the low-skill SEL profile group at T1 and a moderate- or high-skill SEL profile group at T2, they would be classified as having a T1 to T2 *growth* pathway. The resulting SEL growth pathway indicator is then used as an outcome variable in impact models.

Impact Models

Path Impact Model. The Q-ODM tool for describing program impact is called the *path impact model* and can be summarized as shown in Table 2. This simple model is fit to data as a test of the impact of staff instructional quality on children's SEL skill growth, testing hypothesis such as: Exposure to high-quality instructional practices causes more SEL skill growth than exposure to low-quality instructional practices. [21] This research design can be described as a pretest-posttest *multiple groups design* because we examine how the outcome, children's SEL skill growth, varies across several different instructional quality subgroups. [22]

The left-hand column of Table 2 indicates high or low instructional quality, and the top row indicates the three different SEL skill growth pathways. The interior cells represent the results of "crossing" the data for each individual child's quality exposure and skill pathway. The cell entries (i.e., row percentages) represent the proportion of students who evidence each form of SEL skill growth where exposed to each type of quality. Although many comparisons are possible across the cells, the difference in proportions of children in the 'High by Growth' versus 'Low by Growth' cells is of obvious interest and most closely resembles a traditional impact estimate. The extent to which cell proportions differ from what would be expected by chance alone (or from each other) can be estimated using relatively simple statistical tests. [23]

Table 2. Path Impact Model.

		Skill Pathway		
		Growth	**Stability**	**Decline**
Quality	**High**	% of children: high quality & skill growth	% of children: high quality & skill stability	% of children: high quality & skill decline
	Low	% of children: low quality & skill growth	% of children: low quality & skill stability	% of children: low quality & skill decline

The basic impact table can be expanded to include additional quality conditions or pathway patterns. For example, a moderate instructional quality category could added as an additional row to examine whether moderate instructional quality is sufficient for either producing skill growth or preventing skill decline. The pathway columns can also be expanded to include more nuanced pathway patterns. For example, children following a Stability Pathway could be divided into two groups: those who started and ended in the highest SEL skill category versus those who started and ended in a lower SEL skill category.

Prodigal Impact Model. In addition to providing a relatively strong test of the impact of instructional quality on children's SEL skill growth, the *prodigal impact model*[24] can be used to evaluate the equity of how supports and outcomes are distributed across children and settings. We define *equity* as the extent to which settings support SEL skill growth for all children, regardless of their SEL skills at program entry. Equitable programs should promote SEL skill growth for children who enter program settings with low SEL skills and either promote further SEL skill growth or prevent SEL skill decline for children who enter settings with higher SEL skills. Extending from Table 2, Table 3 shows the model for a prodigal impact analysis, which essentially focuses on children who enter program settings with the same baseline SEL skill profiles but whose subsequent SEL skill growth pathways diverge over time (e.g., Stability versus Growth). Equity effects are described by comparing the row percentages for children in cells A and B, with equally high percentages indicating that children who enter high-quality programs with lower SEL skills are receiving benefits similar to children entering high-quality programs with higher SEL skills. Differences in the proportions of children in cells B and C help us understand the potential costs of low instructional quality for children who enter programs with low SEL skills. Both comparisons demonstrate equity effects where results equal or favor the B cell group.

We define equity as the extent to which settings support SEL skill growth for all children, regardless of their SEL skills at program entry. Equitable programs should promote SEL skill growth for children who enter program settings with low SEL skills and either promote further SEL skill growth or prevent SEL skill decline for children who enter settings with higher SEL skills.

Table 3. Prodigal Impact Model.

High Skill at Entry:

		Skill Pathway		
		Growth	**Stability**	**Decline**
Quality	**High**	NA	A	
	Low	NA		

Low Skill at Entry:

		Skill Pathway		
		Growth	**Stability**	**Decline**
Quality	**High**	B		NA
	Low	C		NA

In summary, Table 2 presents the pathway impact model – program quality by child skill pathway – for the MPCn theory of change; it addresses question 4 in Table 1 (i.e., Does exposure to high-quality cause more skill change than exposure to low-quality?). Table 3 extends the discussion of impact to the equity effect, where the quality-outcome interaction is broken out by the low- and high-baseline SEL skill groups, which addresses question 5 in Table 1 (i.e., Do children who were low-skill at baseline gain as much or more than children who were high-skill where exposed to high-quality?). Once the pathway indicators and impact models have been constructed, that information – the pathway indicator or the pathway impact model – can also be examined in relation to other potentially-relevant variables. For example, pathways and impacts can be examined across the levels of any other variable, such as: race/ethnicity categories to model racial equity effects; attendance at 30/60/90 days of program participation to model dosage effects; age groups such as grades K-3, 4-6, 7-12 to model developmental effects. Pathways and impacts can also be examined separately for groups of children across sites as a site-performance metric or aggregated across sites to network or region levels as an aggregate performance metric.

Feedback Tools

Feedback is the most powerful learning support for adults and children (Hattie, 2009). The Q-ODM feedback tools help organizational leaders and evaluators engage stakeholders in data-based decision-making for both CQI and impact evaluation. The feedback tools described here (i.e., Behavior Analogy, Visual Holism, and Benchmarks) heighten the value of feedback by providing readily-actionable information about program quality and its relation to children's SEL skill growth.

The Q-ODM feedback tools help organizational leaders and evaluators engage stakeholders in data-based decision-making for both CQI and impact evaluation.

Behavior Analogy

A pervasive measurement challenge, typically described in terms of *construct validity*, involves the interpretation of scale scores in terms of the specific aspects of the people and contexts about which they are

intended to apply. Although using multiple items to construct scale scores can increase measurement precision, such scale scores can also undermine the power of direct analogy at the lowest level of measurement (i.e., the measurement item) because, unlike items, scale scores tend to lack specificity about object attributes. Q-ODM tools are designed to improve the accuracy and validity of measures, for both the more granular item-level information as well as the information reflected in profile patterns, by promoting the selection and construction of measures suitable for direct analogy to real objects.

Behavioral observation items created using the Q-ODM toolbox generally include granular descriptions of behavior that are direct analogies to the adult and child behaviors that actually occur during OST program offerings. As an example, the top panel of Table 4 contains an observational measurement item and summary data for staff practice aligned to the Awareness part of youths' SEL mental skill, whereas the lower panel contains an observational measurement item and summary data for youths' behavioral expression of the Awareness skill. Both of these items involve the use of awareness skills. During youth planning activities that were happening in less than 30% of programs, only about 50% of youth were clearly using planning strategies.

Although scale scores can be useful for many purposes, we encourage clients to consider the meaning and utility of their data in terms of the specific items used to generate scale scores. For example, although the items shown in Table 4 contribute to their respective scale scores, in which their specific meaning would be lost, there is generally a substantial amount of useful information that can be derived by looking specifically at the item-level information. Accordingly, we generally provide clients with a list of low-scoring items so that they can see the specific staff and youth behaviors they may want to consider or work on. In the current example, such item-level data have some fairly clear implications: First, youth attending these programs are missing opportunities to practice the mental skill of awareness, and many youth are not taking the opportunity when presented. Second, if executive function skills are important to program goals and outcomes, these *infrequently used practices* are clear targets for improvement.

Table 4. Low-scoring Items.

Program Quality Item - Staff provide a structured opportunity for youth to make plans (e.g., youth write down next steps for a project; students converse about how they are going to accomplish goal).	"Practice Not Observed" in 70% of Settings
Youth Behavior Item - How often did the youth evaluate alternative plans for reaching a specific goal (e.g., develop plan-B)?	"Behavior Infrequently Observed" in 44% of children

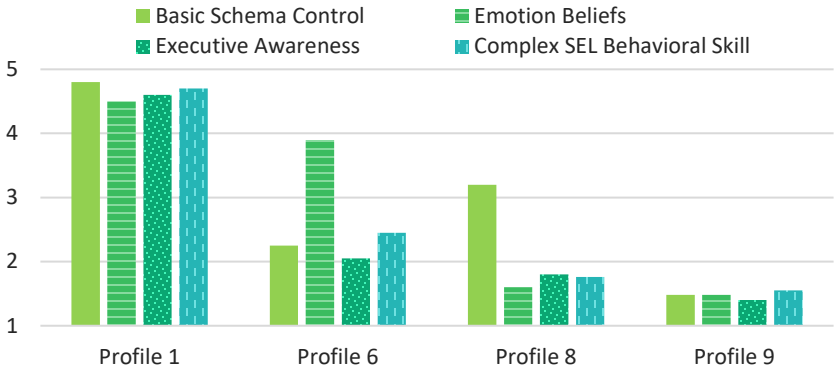
Visual Holism

Holistic *data visualizations* for quality and skill data are useful feedback tools for communicating results and supporting person-in-context reasoning. The meaning of complex multivariate data on, for example, staff practices and children's SEL skills can be simplified by summarizing multiple scale scores as skill and quality profile patterns. The information contained in profile variables is most easily understood when conveyed as a visual-spatial pattern (i.e., the tops of the bar graphs for each profile). Assuming the selected measures have been properly aligned to the relevant attributes of people or settings, the person (child skill) and context (setting quality) profile patterns represent the children and settings as holistic, integrated objects. Further, as described above, the interactions (e.g., impact) between settings and children can then be summarized using relatively simple tables. [25] In essence,

because Q-ODM measurement items and scales are sufficiently detailed, aligned, and complex in theory and composition, their interactions can e modeled relatively simply.

Figure 4 shows an illustrative subset of SEL skill profiles that were identified using four SEL behavioral skill scales, aligned roughly to the four parts of the neuroperson model: Basic *Schema* Control (e.g., behaviorally manages emotion), Emotion *Beliefs* (e.g., uses emotion words), Executive *Awareness* (e.g., makes plans), Complex *Behavior*al Skill (e.g., fulfills group roles). These profiles were drawn from a larger set of nine profiles, so Figure 4 represents about 43% of the sample. [26]

Figure 4. Children's Baseline SEL Skill Profiles.



Consider the two moderately-low profiles: 6 and 8. Profile 6 represents children who entered the OST setting with some knowledge about emotions but who did not manage those emotions well or focus very effectively. Profile 8 represent children who have basic self-control but appear to be withdrawn from engagement with the setting. These two groups – which might be viewed as "externalizers" and "internalizers" – may require responses as different as the those required by the all-high and all-low groups. Also suggested by Figure 4, using either Basic Schema Control (blue bar) or Emotion Beliefs (orange bar) as single variables, it is possible to rate the children in profiles 6 and 8 as having sufficiently high skill for success in the setting when, in fact, they scored low on all of the other parts of the integrated SEL skill set.

This holistic perspective on children's SEL skills highlights the extent to which a group statistics (e.g., means, correlations, regression coefficients) associated with single SEL variables or dimensional constructs may fail to apply to many children or even reveal much useful information about any child. The information provided by the profile pattern (or shape) dramatically improves the interpretability of complex (holistic) information and invites staff into reasoning about real children's SEL skill sets, unclouded by the idiosyncrasies of any particular child and yet with data representing the unique skill set of every individual child in the program. [27]

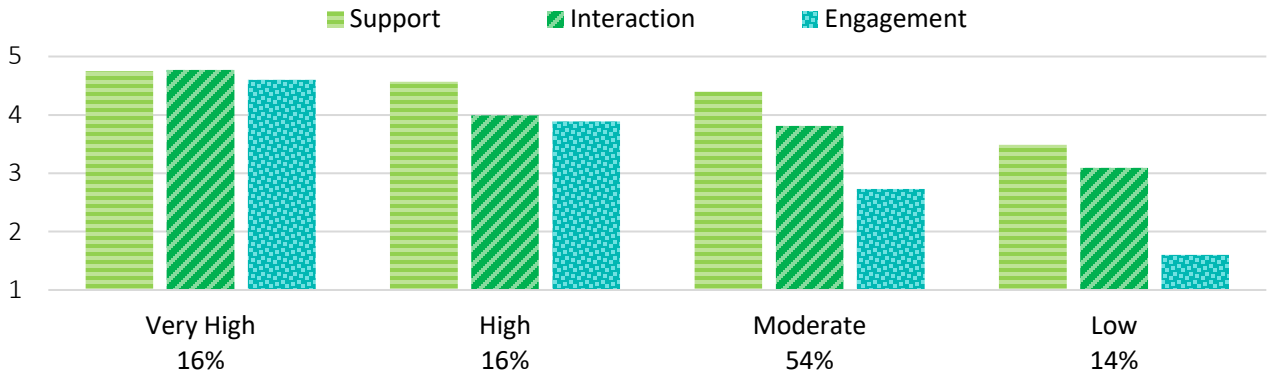
The information provided by the profile pattern (or shape) dramatically improves the interpretability of complex (holistic) information and invites staff into reasoning about real children's SEL skill sets, unclouded by the idiosyncrasies of any particular child and yet with data representing the unique skill set of every individual child in the program… The person (child skill) and context (setting quality) profile patterns represent the children and settings as holistic, integrated objects.

Sites characterized by the highest-quality profile are exemplars of both best practice and programs with strong fit and responsiveness to children's local circumstances and cultures.

Benchmarks

Benchmarks are a critical tool for CQI systems. Information gleaned from prior cycles is typically the most accurate predictor of performance observed during subsequent cycles. Figure 5 presents four instructional quality profiles for OST programs using three SEL quality indicators: Supportive Environment (e.g., staff models skills), Social Interaction (e.g., teams pursue goals), and Executive Engagement (e.g., planning and reflection) from the Palm Beach County Program Quality Assessment (High/Scope, 2006; Lindeman et al., 2019). Each profile represents a subgroup of programs characterized by a *distinct instructional approach*, with 14% of programs characterized by the lowest-performing profile and 16% characterized by the highest-performing profile. In our terminology, benchmarks refer to a pattern of instructional practices (e.g., exemplary or high) that are visible in the profiles and that have also been validated as thresholds at or above which positive impacts on children's SEL skill pathways have generally been observed.

Figure 5. Key Instructional Quality Profile across Three QIS Years.



Drawing from the applied evidence base for this widely used quality assessment tool, we can tell the following story: "Very High" (or, Exemplary) corresponds to very high-quality participatory instruction (aka, positive youth development pedagogy; Smith et al., 2010). Sites characterized by the highest-quality profile are exemplars of both best practice and programs with strong fit and responsiveness to children's local circumstances and cultures. "High" quality instruction is the benchmark for effective OST programs, particularly where serving adolescents, whereas "Moderate" quality instruction corresponds to a blend of middling programs of different types, primarily elementary programs, that are often characterized by the frequent use of direct instruction (aka, staff-centered pedagogy; Smith et al., 2010). "Low" quality instruction corresponds to the inconsistent use of most of the instructional practices associated with each of the instructional quality domains assessed by most versions of the Program Quality Assessment observational rating tool. Settings with low-quality instruction – e.g., where program staff fail to implement basic safety and youth engagement practices – can put vulnerable students at greater risk and compound SEL inequity. Although high-quality instructional practice profiles are the recommended benchmark, exposure to moderate-quality instructional practices may also be sufficient to promote SEL skill growth in some circumstances (Lindeman et al., 2019). Unfortunately, children who enter moderate-quality settings with relatively-high SEL skills often struggle to sustain those skills (i.e., they often evidence SEL skill decline).

III. Evidence for Policy Change

Socio-emotional skill learning can be described as a process of basic and advanced self-regulation that promotes the development of both automatic and intentional forms of agency (Peck & Smith, 2020b). This process is facilitated best where children and youth are supported and trusted to make decisions about things that affect them (Smith, McGovern, Larson, et al., 2016). The Q-ODM tools described here are intended primarily for CQI managers and impact evaluators working in or with OST programs, although they have obvious implications for any educational context (e.g., schools, early childhood). They are designed specifically for use with developmentally-focused interventions; that is, programs focused on SEL skills and academic enrichment through active learning in mixed groupings. Positivist social science methodology has not served well the OST field's needs or purposes and will likely continue the pattern of hegemonic underperformance into the foreseeable future.

In our minds, the key issue is how to most quickly and effectively promote the improvement of services for children, particularly in the face of growing and changing needs that are unlikely to abate. As pandemic, political, and climate-related disruptions to normalcy increase, organizations and institutions will need to be flexible, nimble, and increasingly responsive to such novel conditions. Meaningful performance feedback is perhaps the most integral ingredient in all CQI systems, so we hope that the Q-ODM tools will help organizations become more responsive to their own staff's and children's SEL skill learning needs.

One critical issue that was omitted from our discussion is the cost and practicality of implementing the Q-ODM tools. Here, we simply note that some of these issues have been addressed elsewhere (e.g., Grossman et al., 2009; Smith, 2013; Smith et al., 2018) and that the tools were designed to fit the kinds of data widely collected in the OST field. In addition, it is worth noting that many school districts in the United States, in responding to their state ESSA (Every Student Succeeds Act) plans, are currently collecting most of the different types of data recommended for use with the Q-ODM tools, making pattern-centered approaches to assessing quality-outcome relations for schools primarily a matter of secondary data analysis.

Finally, because the Q-ODM tools include supports for the measurement and modeling of *both* manager/staff practices and children's SEL skills and outcomes, they are especially well suited for evaluating the impact of QIS and other Quality Rating and Improvement Systems (Shen et al., 2011; Zellman et al., 2008). These evaluations require attention to the full cascade of causes and effects that flow from network level rules and resources, to organizational cultures and management, to instructional practices at the point of service, and to children's engagement and SEL skill growth (Lindeman et al., 2019). These improvement systems, and the policies that support them, are important because they offer a viable and evidenced-based way to turn our organizations and services toward greater recognition of the intrinsic, and often untapped, potential that adults and children bring to educational settings.

References

Allport, G. W. (1961). *Pattern and growth in personality.* New York: Holt, Rinehart, & Winston.

Arocha, J. F. (2020). Scientific realism and the issue of variability in behavior. Theory & Psychology, 1-20. doi.org/10.1177%2F0959354320935972

Astbury, B. (2013). Some reflections on Pawson's Science of evaluation: A realist manifesto. *Evaluation: The International Journal of Theory, Research and Practice*, *19*(4), 383–401.

Bailey, J. R., & Eastman, W. N. (1994). Positivism and the promise of the social sciences. *Theory & Psychology*, *4*(4), 505–524.

Bergman, L. R. (1998). A pattern-oriented approach to studying individual development: Snapshots and processes. In R. B. Cairns, L. R. Bergman, & J. Kagan (Eds.), *Methods and models for studying the individual* (pp. 83-121). Thousand Oaks, CA: Sage.

Bergman, L. R., & Magnusson, D. (1991). Stability and change in patterns of extrinsic adjustment problems. In D. Magnusson, L. R. Bergman, G. Rudinger, & B. Törestad (Eds.), *Matching problems and methods in longitudinal research* (pp. 323–346). Cambridge: Cambridge University Press.

Bergman, L. R., Magnusson, D., & El-Khouri, B. M. (2003). *Studying individual development in an interindividual context: A person-oriented approach*. Mahwah, NJ: Erlbaum.

Bergman, L. R., & Vargha, A. (2013). Matching method to problem: A developmental science perspective. *European Journal of Developmental Psychology*, *10*(1), 9–28.

Bornstein, M. H., Hahn, C., & Wolke, D. (2013). Systems and cascades in cognitive development and academic achievement. *Child Development*, *84*(1), 154–162.

Buchanan, D. R. (1998). Beyond positivism: Humanistic perspectives on theory and research in health education. *Health Education Research: Theory & Practice, 13*(3), 439-450.

Bunge, M. (2014). *Chasing reality: Strife over realism*. University of Toronto Press.

Burton-Jones, A., & Lee, A. S. (2017). Thinking about measures and measurement in positivist research: A proposal for refocusing on fundamentals. *Information Systems Research, 28*(3), 451-467.

Cairns, R. B. (1986). Phenomena lost: Issues in the study of development. In J. Valsiner (Ed.), *The individual subject and scientific psychology* (pp. 97-112). New York: Plenum.

Cairns, R. B., & Rodkin, P. C. (1998). Phenomena regained: From configurations to pathways. In R. B. Cairns, L. R. Bergman, & J. Kagan (Eds.), *Methods and models for studying the individual: Essays in honor of Marian Radke-Yarrow* (pp. 245-264). London: Sage.

Chen, H. T., & Rossi, P. H. (1983). Evaluating with sense: The theory-driven approach. *Evaluation Review, 7,* 283-302.

Chirkov, V., & Anderson, J. (2018a). Statistical positivism versus critical scientific realism. A comparison of two paradigms for motivation research: Part 1. A philosophical and empirical analysis of statistical positivism. *Theory & Psychology*, *28*(6), 712–736. doi.org/10.1177/0959354318804670

Chirkov, V., & Anderson, J. (2018b). Statistical positivism versus critical scientific realism. A comparison of two paradigms for motivation research: Part 2. A philosophical and empirical analysis of statistical positivism. *Theory & Psychology*, *28*(6), 737–756. doi.org/10.1177/0959354318816829

Cohen, J. (2006). Social, Emotional, Ethical, and Academic Education: Creating a Climate for Learning, Participation in Democracy, and Well-Being. *Harvard Educational Review*, *76*(2), 201–237.

Cronbach, L. J., & Snow, R. E. (1977). *Aptitudes and instructional methods: A handbook for research on interactions*. New York: Wiley.

Daly, M. (1973). *Beyond God the Father*. London: Women's Press.

Danielson, C. (1996). *Enhancing professional practice: A framework for teaching*. Alexandria, Virginia: Association fo rsupervision and curriculum development.

Danziger, K. (1990). *Constructing the subject: Historical origins of psychological research*. Cambridge University Press.

Diamantopoulos, A., Reifler, P., & Roth, K. P. (2008). Advancing formative measurement models. *Journal of Business Research* (61:12), 1203-1218.
Dynarski, M., James-Burdumy, S., Mansfield, W., Mayer, D., Moore, M., Mullens, J., Silva, T., Pistorino, C., & Hermond, D. (2001). A Broader View: The National Evaluation of the 21st Century Community Learning Centers Program - Design Report: Volume I. Princeton, NJ: Mathematica Policy Research.
Eccles, J. S., & Gootman, J. A. (2002). *Community programs to promote youth development*. Washington, DC: National Academy Press.
Eronen, M. I. (2019). Robust realism for the life sciences. *Synthese*, *196*, 2341–2354.
Feinstein, L., & Peck, S. C. (2008). Unexpected educational pathways: Why do some students not succeed in school and what helps others beat the odds? *Journal of Social Issues, 64*, 1–20.
Grice, J. W. (2004). Bridging the idiographic-nomothetic divide in ratings of self and others on the big five. *Journal of Personality*, *72*(2), 203–241.
Grice, J. W. (2015). From means and variances to persons and patterns. *Frontiers in Psychology*, *6*. dx.doi.org/10.3389/fpsyg.2015.01007
Grice, J. W., Jackson, B. J., & McDaniel, B. L. (2006). Bridging the idiographic-nomothetic divide: A follow-up study. *Journal of Personality*, *74*(4), 1191–1218.
Grice, J. W., Yepez, M., Wilson, N. L., & Shoda, Y. (2017). Observation-oriented modeling: Going beyond "Is it all a matter of chance"? *Educational and Psychological Measurement*, *77*(5), 855–867.
Grossman, J. B., Lind, C., Hayes, C., McMaken, J., & Gersick, A. (2009). *The cost of quality of out-of-school time programs.* New York: Public/Private Ventures.
Haig, B. D., & Evers, C. W. (2015). *Realist inquiry in social science*. Sage.
Held, B. S. (2014). Realism, reification, and monism. *Journal of Theoretical and Philosophical Psychology*, *34*(3), 187–194.
Harms, T., & Clifford, R. (1980). Early Childhood Environment Rating Scale (ECERS). New York: Teachers College Press.
Hattie, J. (2009). *Visible learning: a synthesis of over 800 meta-analyses relating to achievement*. New York: Rutledge.
Heckman, J. J. (2007). The economics, technology, and neuroscience of human capability formation. *Proceedings of the National Academy of Sciences, 104*(33), 13250-13255.
High/Scope (2003). *PQA preschool program quality assessment*. Ypsilanti, MI: High/Scope Educational Research Foundation. (https://highscope.org/)
High/Scope (2006). *Palm Beach County Program Quality Assessment.* Ypsilanti, MI: High/Scope Educational Research Foundation. (https://highscope.org/)
Ho, A. D. (2008). The problem with "proficiency": Limitations of statistics and policy under No Child Left Behind. *Educational Researcher, 37*(6), 351–360.
Holland, J. H. (1995). *Hidden order: How adaptation builds complexity.* New York: Addison-Wesley.
Husband, T., & Hunt, C. (2015). A review of the empirical literature on No Child Left Behind from 2001 to 2010. *Planning and Changing, 46*(1/2), 212-254.
Karoly, L. A. (2014). *Validation Studies for Early Learning and Care Quality Rating and Improvement Systems: A Review of the Literature*. www.rand.org/pubs/working_papers/WR1051.html
Kluckhohn, C., & Murray, H. A. (1948). Personality formation: The determinants. In C. Kluchhohn & H. A. Murray (Eds.), *Personality in nature, society, and culture* (pp. 35-48). New York: Knopf.
Ladd, H. F. (2017). No Child Left Behind: A deeply flawed federal policy. *Journal of Policy Analysis and Management, 36*(2), 461-469.
Lamiell, J. T. (2013). Statisticism in personality psychologists' use of trait constructs: What is it? How was it contracted? Is there a cure? *New Ideas in Psychology*, *31*(1), 65–71.
Larson, R. (2000). Toward a psychology of positive youth development. *American Psychologist, 55*, 170-183.

Leplin, J. (1986). Methodological realism and scientific rationality. *Philosophy of Science, 53*(1), 31-51.
Lester, P. (2018). *Evidence-based comprehensive school improvement: How using proven models and practices could overcome decades of failure*. White Paper. Social Innovation Research Center. http://www.socialinnovationcenter.org/
Lindeman, L. M., Smith, C., Peck S. C., & Harvey, S. L. (2019). *Evidence of positive impact on afterschool and summer programs in Palm Beach County: A 10-year validation study of Prime Time's integrated quality improvement system*. Boynton Beach, FL: Prime Time Palm Beach County. https://simplebooklet.com/10yearvalidationstudy
Mackenzie, B. (1977). *Behaviourism and the limits of scientific method.* Atlantic Highlands, NJ: Humanities.
Magnusson, D. (2003). The person approach: Concepts, measurement models, and research strategy. In W. Damon (Series Ed.), S. C. Peck, & R. W. Roeser (Vol. Eds.), New *Directions for Child and Adolescent Development: Vol. 101. Person-centered approaches to studying human development in context* (pp. 3-24). San Francisco: Jossey-Bass.
Masten, A. S., & Cicchetti, D. (2010). Developmental cascades. *Development and Psychopathology*, *22*(3), 491–495.
McGill, E., Er, V., Penney, T., Egan, M., White, M., Meier, P., Whitehead, M., Lock, K., Anderson de Cuevas, R., Smith, R., Savona, N., Rutter, H., Marks, D., de Vocht, F., Cummins, S., Popay, J., & Petticrew, M. (2021). Evaluation of public health interventions from a complex systems perspective: A research methods review. *Social Science & Medicine*, 272, 1-18. doi.org/10.1016/j.socscimed.2021.113697
Michell, J. (2003). Pragmatism, positivism and the quantitative imperative. *Theory & Psychology*, *13*(1), 45–52.
Pavitt, C. (2016). Realism. *The International Encyclopedia of Communication Theory and Philosophy*. DOI: 10.1002/9781118766804.wbiect005
Pawson, R., & Tilley, N. (1997). *Realistic evaluation*. Sage.
Peck, S. C. (2007). TEMPEST in a gallimaufry: Applying multilevel systems theory to person-in-context research. *Journal of Personality, 75*, 1127-1156
Peck, S. C. (2009). Using multilevel systems theory to integrate dynamic person-in-context systems. *Journal of Research in Personality, 43,* 262-263.
Peck, S. C., Roeser, R. W., Zarrett, N., & Eccles, J. S. (2008). Exploring the roles of extracurricular activity quantity and quality in the educational resilience of vulnerable adolescents: Variable- and pattern-centered approaches. *Journal of Social Issues, 64*, 135-155.
Peck, S. C., & Smith, C. (2020a). *Impact evaluation for the* Parent Child Plus *program, Newark Trust for Education.* Ypsilanti, MI: QTurn.
Peck, S. C., & Smith, C., (2020b). *Socio-Emotional Skills, Quality, and Equity: The Multilevel Person-in-Context~neuroperson (MPCn) Framework.* [White Paper #1]. QTurn. https://qturngroup.com/knowledgecenter/white-paper-1/
Pepper, S. (1942) *World hypotheses*. Berkeley: University of California Press.
Pianta, R., & Hamre, B. (2009). Conceptualization, measurement, and improvement of classroom processes: Standardized observation can leverage capacity. *Educational Researcher, 38*, 109-119.
Richters, J. E. (1997). The Hubble hypothesis and the developmentalist's dilemma. *Development and Psychopathology, 9,* 193-229.
Rihoux, B., & Ragin, C. (2008). *Configurational comparative methods: Qualitative comparative analysis (QCA) and related techniques*. Thousand Oaks, CA: Sage.
Roeser, R. W., & Peck, S. C. (2003). Patterns and pathways of educational achievement across adolescence: A holistic-developmental perspective. In W. Damon (Series Ed.) & S. C. Peck & R. W.

Roeser (Vol. Eds.), *New Directions for Child and Adolescent Development: Vol. 101. Person-centered approaches to studying human development in context* (pp. 39-62). San Francisco: Jossey-Bass.
Roeser, R. W., & Peck, S. C. (2009). An education in awareness: Self, motivation, and self-regulated learning in contemplative perspective. *Educational Psychologist, 44*, 119-136.
Roeser, R. W., Peck, S. C., & Nasir, N. S. (2006). Self and identity processes in school motivation, learning, and achievement. In P. Alexander & P. H. Winne (Eds.), *Handbook of educational psychology* (2nd edition) (pp. 391-424). Mahwah, NJ: Erlbaum. Rogers et al., 2000
Rose, T., Rouhani, P., & Fischer, K. W. (2012). The Science of the Individual. *Mind, Brain, and Education*, *7*(3), 152-158.
Schneider, C. Q., & Wagemann, C. (2012). *Set-theoretic methods for the social sciences*. Cambridge: Cambridge University Press.
Shen, J., Ma, X., Tackett, W. L., Lu, X., Brandi, K., Goodman, J., Till, L., & Watson, G. (2011). Effects of quality improvement system on early childhood education practitioners. In J. A. Sutterby (Ed.), The Early Childhood Educator Professional Development Grant: Research and Practice, Vol. 15 (pp. 129-154). Emerald Group. https://doi.org/10.1108/S0270-4021(2011)0000015009
Silver Strong & Associates (2013). The Thoughtful Classroom Teacher Effectiveness Framework. https://thoughtfulclassroom.com/
Smith, C. (2013). Moving the needle on moving the needle: Next stage technical guidance for performance based accountability systems in the expanded learning field with a focus on performance levels for the quality of instructional services. . Washington DC: David P. Weikart Center for Youth Program Quality at the Forum for Youth Investment.
Smith, C., Akiva, T., Sugar, S. A., Lo, Y. J., Frank, K. A., Peck, S. C., Cortina, K., & Devaney, T. (2012). *Continuous quality improvement in afterschool settings: Impact findings from the Youth Program Quality Intervention study*. Washington, DC: Forum for Youth Investment.
Smith, C., & Hohmann, C. (2005). Full findings from the Youth PQA validation study. Ypsilanti, MI: *High/Scope Educational Research Foundation.*
Smith, C., McGovern, G., Larson, R. W., Hillaker, B., & Peck, S. C. (2016). *Preparing youth to thrive: Promising practices for social emotional learning*. Washington, DC: David P. Weikart Center for Youth Program Quality at the Forum for Youth Investment.
Smith, C., McGovern, G., Peck., S.C., Larson, R., Hillaker, B., Roy, L. (2016). *Preparing youth to thrive: Methodology and findings from the social and emotional learning challenge.* Washington DC: David P. Weikart Center for Youth Program Quality at the Forum for Youth Investment.
Smith, C., & Peck S. C. (2019). *Impact evaluation for Palm Beach County quality improvement system (QIS) using fully pattern-centered analytics*. Ypsilanti, MI: QTurn.
Smith, C., & Peck, S. C. (2020a). *Measuring socio-emotional skill and equity in out-of-school time settings* [White Paper #2]. QTurn. https://qturngroup.com/knowledgecenter/white-paper-2/
Smith, C., & Peck, S. C. (2020b). *Citizen Science and Advocacy in OST* [White Paper #4]. QTurn. https://qturngroup.com/knowledgecenter/white-paper-4/
Smith, C., Peck, S. C., Denault, A. S., Blazevski, J., & Akiva, T. (2010). Quality at the point of service: Profiles of practice in after-school settings. *American Journal of Community Psychology, 45,* 358-369.
Smith, C., Peck, S. C., Macleod, C., Roy, L, Helegda, K, & Borah, P. (2018). Realizing the Potential of Summer Learning: Design and Development Study for the Summer Learning Program Quality Intervention. White Paper. Ypsilanti, MI: The David P. Weikart Center for Youth Program Quality - A Division of the Forum for Youth Investment.
Smith, C., Peck, S. C., & McNeil, B. (2019a). *Promoting healthy development of young people: Outcomes Framework 2.0*. London, UK: Centre for Youth Impact.

Smith, C., Peck, S., Pittman, K., McGovern, G. (2015). Framing an evidence-based decision about 21st CCLCL: How do we see the value. Policy Commentary. Washington DC: Forum for Youth Investment.

Smith, C., Peck, S. C., Roy, L., & Smith, L. (2019b, April). Measure once, cut twice: Using data for continuous improvement and impact evaluation in education programs. In D. Darling (Chair), *Digging deeper: Using mixed-methods research to assess promising practices in out-of-school time*. Symposium conducted at the annual meeting of the American Educational Research Association, Toronto, ON.

Smith, C., Ramaswamy, R., Helegda, K., Macleod, C. et al. (2017) Design study for the Summer Learning Program Quality Intervention (SLPQI): Final-year intervention design and evaluation results. Washington DC: David P. Weikart Center for Youth Program Quality at the Forum for Youth Investment.

Smith, C., Roy, L., Peck, S. C., & Macleod, C. (2017). Evaluation of Program Quality and Social and Emotional Learning in American Youth Circus Organization Social Circus Programs. Ypsilanti, MI: The David P. Weikart Center for Youth Program Quality - A Division of the Forum for Youth Investment.

Smith, C., Roy, L., Peck, S., Macleod, C., Helegda, K. (2017). Quality-outcomes study for Seattle Public Schools summer programs, 2016 program cycle. Washington DC: David P. Weikart Center for Youth Program Quality at the Forum for Youth Investment.

Staats, A. W. (1991). Unified positivism and unification psychology: Fad or new field? *American Psychologist*, *46*(9), 899–912.

Stame, N. (2004). Theory-based evaluation and varieties of complexity. *Evaluation*, *10*(1), 58-76.

Stedman, J. M., Kostelecky, M., Spalding, T. L., & Gagné, C. (2016). Scientific realism, psychological realism, and Aristotelian–Thomistic realism. *Journal of Mind and Behavior*, *37*(3–4), 199–218.

Stern, E., Stame, N., Mayne, J., Forss, K., Davies, R., & Befani, B. (2012). Broadening the range of designs and methods for impact evaluations, DFID Working Paper 38, London: Department for International Development. http://r4d.dfid.gov.uk/Output/189575/

Thornburg, K. R., Mayfield, W. A., Hawks, J. S., & Fuger, K. L. (2009). The Missouri quality rating system school readiness study. Columbia, MO: Center for Family Policy & Research.

Uher, J. (2019). Data generation methods across the empirical sciences: Differences in the study phenomena's accessibility and the processes of data encoding. *Quality & Quantity: International Journal of Methodology*, *53*(1), 221–246.

Uher, J. (2021). Psychometrics is not measurement: Unraveling a fundamental misconception in quantitative psychology and the complex network of its underlying fallacies. *Journal of Theoretical and Philosophical Psychology*, *41*(1), 58–84.

University of Washington Center for Educational Leadership (2016). The 5 Dimensions of Teaching and Learning. https://k-12leadership.org/

Urban, J. B., & Trochim, W. (2009). The role of evaluation in research-practice integration: Working toward the "golden spike." *American Journal of Evaluation*, *30*(4), 538–553.

Vargha, A., Torma, B., & Bergman, L. R. (2015). ROPstat: A general statistical package useful for conducting person-oriented analyses. *Journal for Person-Oriented Research, 1*, 87-98.

von Eye, A., & Bergman, L. R. (2003). Research strategies in developmental psychopathology: Dimensional identity and the person-oriented approach. *Development and Psychopathology, 15*, 553-580.

von Eye, A., & Niedermeier, K. E. (1999). *Statistical analysis of longitudinal categorical data in the social and behavioral sciences: An introduction with computer illustrations.* Mahwah, NJ: Erlbaum.

Zellman, G. , Perlman, M., Le, V., & Setodji, C. M. (2008). Assessing the validity of the Qualistar early learning quality rating and improvement system as a tool for improving child-care quality. Santa Monica, CA: RAND. https://www.rand.org/

Znaniecki, F. (1934). *The method of sociology*. New York: Rinehart.

Notes

[1] We refer to variable-centered theory and methods, psychometrics, and generalized linear modeling strategies in particular (e.g., ANOVA, regression), as *positivist* theory and methodology. As described within the disciplines of history and philosophy of science (Arocha, 2020; Baily & Eastman, 1994; Buchanan, 1998; Burton-Jones & Lee, 2017; Chirkov & Anderson, 2018a, 2018b; Michell, 2003; Staats, 1991), application of positivist methods results in pacification of the research participants (e.g., staff, students) by limiting the range of relevant performance indicators to behavior and simple beliefs (Danziger, 1990; Mackenzie, 1977), in part because of the tight linkage between positivist methodology and the psychological ideology of behaviorism that, in its modernized form, also emphasizes a reductionist perspective (in this case *operationism*) on mental skills. One of the primary benefits of realist and multilevel thinking is that data collection, reduction, and analysis are more likely to reflect a match between theory and method. If we accept that person-in-context systems are real, multilevel systems, then we should also accept that traditional forms of positivist, variable-centered thinking and analysis are severely limited (Bergman & Vargha, 2013; Bergman et al., 2003; Cairns, 1986; Holland, 1995; von Eye & Bergman, 2003).

[2] For simplicity, we generally use the term *children* where referring to children and youth, ages 5 to 18.

[3] Behaviorist theory is generally incompatible with person-centered theory, mainly because it ignores mental skills and treats behavioral information as if it were sufficient for understanding children's development (Mackenzie, 1977). Psychometrics are used to engender a false sense of precision and generally fail to accurately represent either the different parts of an individual's SEL skill set or the integrated SEL skill set as a whole (Chirkov & Anderson, 2018a, b; Grice et al., 2017; Uher, 2019). Linear statistical models assume that people (and contexts) are homogeneous (Richters, 1997) and focus on average effects that do not necessarily apply to any particular person (or context) (Rose et al., 2012; von Eye & Bergman, 2003), rendering them theoretically "anemic" (Grice et al., 2017) and compounding the problems of psychometrics and intent-to-treat experimentalism. The counterfactual contrast (particularly of the no-treatment variety) lacks ecological validity and obscures meaningful individual differences in SEL history, SEL skill change, and exposure to various degrees and forms of treatment.

[4] Although there are many varieties of realism (cf. Astbury, 2013; Eronen, 2019; Stedman et al., 2016), according to Arocha (2020), "all scientific realists adopt at least two basic theses: one ontological and one epistemological (Bunge, 2014; Haig & Evers, 2015). The ontological thesis is that the world is real and it exists independently of our knowledge of it. The epistemological thesis is that reality is knowable, albeit often approximately and mostly indirectly" (p. 3). This is the kind of *realist(ic)* approach we apply to CQI and impact evaluation.

[5] In general, the pattern-centered theories and methods in the Q-ODM toolbox were developed by social scientists interested in methodology and holistic, developmental systems theories characterized by concepts such as self-organization, process dynamics, person-by-context interactions, ecological and social contexts, multilevel systems, and life-course or lifespan development. Much of this work was in direct response to the perceived limitations of narrowly-defined behaviorist theory, psychometrics, variable-centered statistical models (e.g., ANOVA and regression), and the overly strict reliance on experimental methods and counterfactual designs (i.e., the positivist approach to social science). Pattern-centered approaches provide a realist, and realistic, alternative to these positivist measurement and analytic approaches.

[6] We use the term methodological realism (Leplin, 1986) to indicate that the methods we use to measure and model the constructs included in our theoretical framework (e.g., schemas, beliefs, and awareness) are based on the assumptions that (a) the construct names refer to real entities (e.g., not conceptual abstractions) and (b) properly-aligned methods can yield information about the nature and function of these entities.

[7] We use the term psychological realism (Pavitt, 2016) to indicate our assumption that the constructs included in our theoretical framework (e.g., schemas, beliefs, and awareness) *are* something real and *do* something real (cf. Allport, 1961). However, although we tend to assume "the *nonepiphenomenal* status of psychological kinds, that is, their causal powers, or their participation in causal relations" (Held, 2014, p. 188), we qualify this assumption by noting, for example, that any given thought or feeling may or may not be nonepiphenomenal, depending on its

specific role in specific instances. In other words, a thought might be epiphenomenal in relation to contemporaneous behavior while being nonepiphenomenal in relation to the contemporaneous construction and encoding of a new belief.

[8] There are variety of ways to examine program impacts that do not involve randomized-control trial (RCT) designs. For example, consistent with theory-based approaches to evaluation that emphasize realist conceptualizations of substantive and cascading causal forces (Chen & Rossi, 1983; Pawson & Tilley, 1997; Rogers et al., 2000; Stame, 2004; Stern et al., 2012; Urban & Trochim, 2009), quasi-experimental designs applicable to most OST programs involve no random assignment of participants or settings to a "control" or comparison group that does not experience exposure to the program. Our pattern-centered approach to impact evaluation can include RCT designs but is more akin to methods designed for understanding *complex adaptive systems* (Holland, 1995) than simple bivariate relations between the 'intent to treat' and the intended program outcomes (e.g., Rihoux & Ragin, 2008; Schneider & Wagemann, 2012; Stern et al., 2012).

[9] Staff reports on management practices are generally viewed as more valid than manager self-reports of their own practices due to several common sources of survey method bias, such as *demand characteristics* (e.g., where respondents provide answers that fit their expectations about the purpose of the study or measure) and *social desirability* (e.g., where respondents provide answers that make them look good or avoid embarrassment).

[10] In direct response to adult practice, youth engagement occurs in each moment of OST programming. Measures of mental engagement correspond to the Awareness part in Figure 2; specifically, the thoughts and feelings young people experience while at the point of service. Our recommended measure of youth engagement includes youth self-reports of interest, challenge, and stress during a specific offering.

[11] Program quality and setting measures include the Youth Program Quality Assessment (Smith & Hohmann, 2005), the Classroom Assessment Scoring System (Pianta & Hamre, 2009), the Early Childhood Environment Rating Scale (Harms & Clifford, 1980), the Preschool Program Quality Assessment (High/Scope, 2003), and numerous other 'teacher' observation tools now on the market, such as the Danielson Framework (Danielson, 1996), the 5 Dimensions of Teaching and Learning (University of Washington Center for Educational Leadership, 2016), and the Thoughtful Classroom Teacher Effectiveness Framework (Silver Strong & Associates, 2013).

[12] Further, cause-effect relations along such *developmental cascades* (Bornstein et al., 2013; Masten & Cicchetti, 2010) are much stronger (and often strictly) between objects at adjacent levels than more distal levels. For example, high-quality instructional practices are more strongly associated with children's engagement at the point of service and SEL skill change than with transfer outcomes.

[13] For readers familiar with the literature on the aptitude-treatment interaction (Cronbach & Snow, 1977), this group of questions has been of great interest to psychologists and education evaluators within the positivist tradition, even though their efforts to successfully address these questions have been undermined substantially by inconsistencies of theory, method, and results.

[14] Reconfiguring items into scales not intended by test developers can be a useful or necessary way to address validity issues that have direct practical utility by making previously conceived and collected data more useful for everything that follows (e.g., subsequent manipulations, interpretations, and decisions).

[15] If observers can reliably identify particular behaviors in a specific types of settings, different aspects of behavior can be defined as indicators of high or low skill by being present or absent for any child in a parallel setting. However, this kind of indicator often requires a methodological grounding in formative measurement models (Diamantopoulos et al., 2008) and modeling with ordinal and categorical variables (Grice, 2015).

[16] If each variable selected for inclusion in a holistic description of a person or context has the same set of response options, and these response options have clear interpretations after combining items into scales, these variables can generally be used in their original form. However, the process of combining items into scales often involves first converting the original response scale values across a set of items that use different response scale values to a common *quasi-absolute* response scale (Bergman & Magnusson, 1991; Bergman et al., 2003). Ideally, this standardization is based on expert practitioner knowledge about critical thresholds in the respective scales.

[17] As an example, consider the following behavioral observation rating Item and corresponding response scale (i.e., 0 = never, 1 = about half the time, 2 = almost all the time): *How often did the youth describe the reason for their feelings when upset (e.g., "I got into a fight with my mom" or "he called me names")?* During quasi-absolute scaling, local expert practitioners told us that a score of 2 was the meaningful threshold for self-regulation in their programs: children scoring in the 1 or lower range were in need of more support from staff and were not likely to be experiencing positive youth development. In contrast, for the item, *How often did the youth easily manage both positive and negative feelings (e.g., didn't lash out at others when feeling bad; didn't brag or gloat when feeling good)?* using the same response scale, expert practitioners said that a score of 2 was the benchmark for effectiveness in the setting.

[18] For example, the pattern-centered measurement approach yields categorical variables indicating a specific SEL skill profile for each child at each point in time (see White Paper 2, Smith & Peck, 2020a), and the pattern-centered analytic approach yields intraindividual child SEL skill growth pathways by linking time-specific SEL skill profiles across time (e.g., using LICUR – i.e., linking clusters after removal of a residue [Bergman, 1998] – cross-tabulation analysis, or loglinear modeling [von Eye & Niedermeier, 1999]). This method identifies normative patterns of child SEL skill growth (or instructional practices), where they exist, yet also allows for a mid-range of generalizability reflecting subgroups of children (or adults) characterized by relatively non-normative profiles and pathways of SEL skill growth (or instructional quality) that are difficult to represent using linear quantitative models (because the scores from such non-normative cases are typically handled, by default, as if they reflect measurement or modeling error in variable-centered models).

[19] In contrast to most variable-centered methods, which define growth in terms of rank-order stability (e.g., whether a child grows more or less than another child, as opposed to the absolute amount of growth) or mean-level stability (which reflects the average shape of growth) on single variables, pattern-centered methods focus on holistic, person- or context-centered, multivariate growth pathways. Multivariate growth pathways are defined by first identifying multivariate profiles for each person (or context) at two or more points in time, ideally where all variables have response-scale values with the same meaning at every point in time, and then by examining how each person's (or context's) profile shape changes as they move through time (e.g., moving from lower-skill to higher-skill profiles). This approach to assessing children's SEL skill growth, as well as its causes and consequences, requires first identifying the range of sample-level *growth states* (i.e., time-specific profiles) into and out of which children move through time, before considering each child's individual-level growth pathway. For example, the profile pattern for the highest profile at the first timepoint might correspond more closely to the profile pattern for the second-highest profile at the second time point; in this case, the highest profile at time 1 would be coded "2" (instead of "1") so that it corresponds to the second-highest profile at the second time point, which would also be code "2." Following this alignment across time, individual-level pathways can be coded into growth, stability, and decline categories by reference to the sample-level profile patterns that were aligned across all timepoints.

[20] The pattern-centered model for individual-level skill growth (e.g., student or staff skill) requires first calculating and aligning profile patterns by reference to the profile *centroids* - the geometric center of the variable values corresponding to a profile pattern. The distance between two profile centroids (i.e., the set of means on the profile variables within each profile group) can be described in terms of the average squared Euclidean distance (ASED). The ASED metric is used to generate quantitative estimates of the similarity between profile patterns across different timepoints. With this quantitative similarity information, we can align the sample-level profile patterns across time so that, for example, the meaning of a "high-quality" profile pattern is the same at all points in time (see note 19, above, and Appendix D, in Peck & Smith, 2020a).

[21] The logic of this path impact analysis can be extended by integrating profile and path variables for both quality and SEL skills into variable-centered models, along with any potentially relevant confounding variables (e.g., adding covariates or propensity scores). Including these covariates in the model provides a stronger basis for establishing the extent to which SEL skill changes are caused by differences in instructional quality as opposed to other factors, like selection bias, in which 'pre-test' differences on potentially confounding variables between children exposed to high- versus low-quality cause the observed pathway differences.

[22] This "skill growth by levels of quality" design has been used with some frequency in early childhood evaluations (e.g., Karoly, 2014; Thornburg et al., 2009) and was the subject of extensive study in the literature on aptitude-

treatment interactions (Cronbach & Snow, 1977). In our analysis, the main threats to the validity of a conclusion such as, higher-quality instructional practices have more positive impacts on children's SEL skill growth than lower-quality instructional practices, would be *selection bias*, or pre-existing differences between children exposed to high- versus low-quality instruction, or *selection-maturation* interaction, where the pre-existing differences pertain specifically to differences in children's SEL skill maturation rates. Each impact effect can be examined separately at two or more levels of a potentially-confounding variable; or, profile and pathway variables can be integrated into standard linear models (e.g., logistic regression models that use student SEL growth pathway variables as the criterion, staff instructional profile variables as predictors, and potentially-confounding variable as covariates; Peck et al., 2008).

[23] An omnibus chi-square test reveals the extent to which there are systematic relations between instructional quality and skill growth. Cell-specific adjusted standardized residuals reveal the extent to which each of the observed cell counts differ from what would be expected from chance relations between each of the respective forms of instructional quality and skill growth.

[24] Where there are no measures of potentially-confounding variables or no information about what kinds of potentially-confounding variables might exist, impact effects can be examined using *prodigal* analysis (Cairns & Rodkin, 1998). Prodigal analyses shift the focus from the full range of possible pathways (e.g., of SEL skill growth) to a series of focused contrasts between (a) children who follow the pathway that would be normatively expected given their initial profile pattern (e.g., low-skill children remaining low-skill across time) and (b) children who deviate from the pathway that would be normatively expected given their initial profile pattern (e.g., low-skill children who develop higher skills across time). By framing these prodigal analyses in terms of all the children who begin a particular time period in the same profile subgroup, we guard against potential selection effects by increasing the probability that children who share the same baseline profile also share characteristics that may otherwise differentially impact their growth or, in particular, the effects of staff instructional quality on their SEL skill growth (Roeser & Peck, 2003; Feinstein & Peck, 2008).

[25] Although "it seems reasonable to assume that because developmental phenomena are multi-determined and complex, the procedures and statistical analyses employed to study them should be equally complex…. Exactly the opposite conclusion may be reached if the principal task for developmental research is to understand and clarify…. Parsimony in analysis may be permitted because the major analytic solutions have already been reached in the conception of the study, in the *methods adopted*, and in the *forms of the data available* for analysis (Cairns, 1986, p. 100, emphasis added).

[26] These data were drawn from a baseline sample of 1159 mostly elementary-aged children in OST programs in Palm Beach County, Florida during 2017.

[27] Organizing multilevel, multivariate complexity by reference to relatively-homogeneous subgroups allows us to avoid the pitfalls associated with overly general sample-level averages and overly specific idiosyncratic particulars (cf. Kluckhohn & Murray; 1948; Roeser & Peck, 2003). In short, dispensing with typically unrealistic assumptions characterizing variable-centered approaches (e.g., linear, additive interactions across variables and homogeneity in casual structures across persons and contexts) allows us to represent complex interactions characterizing proximally-integrated systems with fairly simple, pattern-centered, categorical variables (cf. Grice, 2004; Grice et al., 2006, 2015).