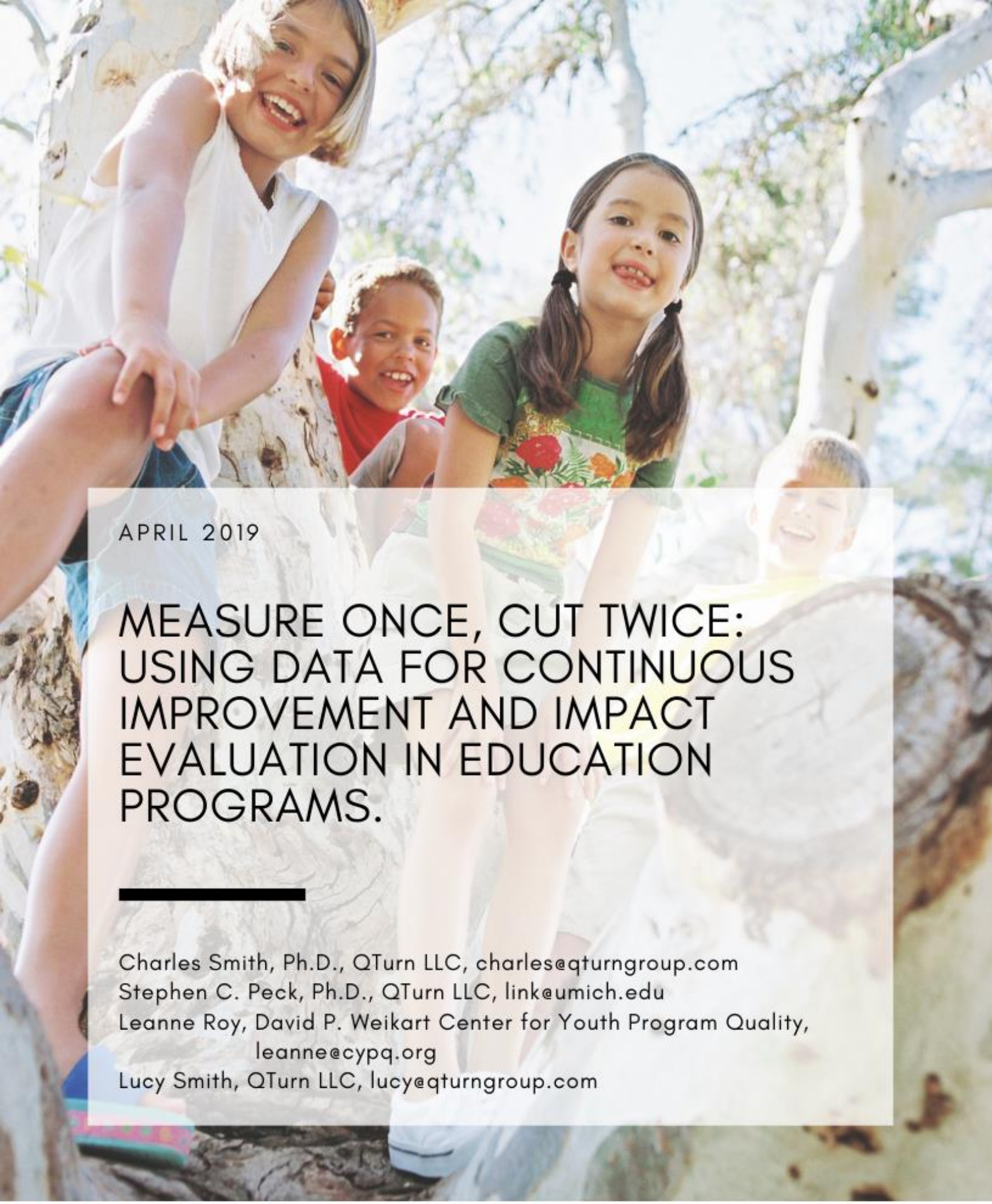


APRIL 2019

MEASURE ONCE, CUT TWICE: USING DATA FOR CONTINUOUS IMPROVEMENT AND IMPACT EVALUATION IN EDUCATION PROGRAMS.

Charles Smith, Ph.D., QTurn LLC, charles@qturngroup.com
Stephen C. Peck, Ph.D., QTurn LLC, link@umich.edu
Leanne Roy, David P. Weikart Center for Youth Program Quality, leanne@cypq.org
Lucy Smith, QTurn LLC, lucy@qturngroup.com

QTURN

Contents

Citation: Smith, C., Peck S., Roy, R., Smith, L. (2019). Measure once, cut twice: Using data for continuous improvement and impact evaluation in education programs. Meetings of the American Education Research Association, Toronto, ON, CA.

Introduction

Frustration and confusion often occur when practitioners require detailed information about program processes for continuous quality improvement (CQI) while policy-makers require evidence of outcome effects for accountability and funding. Impact studies are often preferred over continuous improvement studies, but they seldom offer useful information to practitioners. Per the conference theme, this situation leads to a worldview that emphasizes the limitations of social science methods for achieving practical purposes and welcomes arbitrary decision making (i.e., Type-2 error) in the absence of better evidence and arguments.

As an example, the James-Burdumy et al. (2005) federally commissioned evaluation of the 21st Century Community Learning Centers program is frequently cited as evidence – from a set of rigorous quasi-experimental tests – that afterschool programs are ineffective. However, it is also likely that some outcome effects were missed because the study did not include adequate measurement and modeling of program implementation and other proximal indicators of instructional quality, student skill learning needs, and student skill change. In addition to the overall pattern of nil effects on student outcomes, it is also possible that (a) program or instructional quality was too low to produce an effect on student outcomes in a large proportion of the sample, masking real effects that occurred in a smaller number of programs or (b) real skill gains occurred in different skill domains (e.g., emotion management, empathy, mathematics, ecology) for small subgroups of students such that focusing on the average effect of program participation on growth of any single skill masked simultaneous effects in different domains (Smith, Peck, Pittman, McGovern, 2015).

The fact that implementation, instructional quality, and proximal skill growth were not major areas of focus in such an important program evaluation points clearly to larger issues for education policy: Implementation failure is likely to be both prevalent and unidentified in direct service organizations that provide a wide array of education and human services. As a corollary, implementation successes – e.g., best practice, high fidelity, high quality, meeting standards – are also likely prevalent and unidentified. In this circumstance, students cannot be protected from exposure to low quality, staff cannot be rewarded for producing high quality, and managers cannot make effective decisions about how to use resources.

Although education policies that include requirements and resources for *both* CQI and impact evaluation are increasing in number, several specific challenges (discussed below) have slowed progress. We need policies that (a) better blend the sciences of *program evaluation* and

performance measurement in ways that improve methodology and lower costs (McDavid, Huse, & Hawthorn, 2013) and (b) result in tangible improvements in the *quality of services* to levels at which outcome effects are known to occur. This paper describes a generic *quality-outcomes design* (Q-O design) that meets the need for *performance measurement methodology* for concurrent and integrated impact evaluation and continuous improvement in the same organization; that is, *measure once, cut twice*.

Review

Traditional definitions of rigor entailed in psychometrics and experimental design often put these purposes – impact and improvement– on different sides of a methodological fence. Counterfactual reasoning suggests that process quality and proximal skill assessment[i] are often unnecessary and always unreliable, and that these issues, particularly with observational ratings, are expensive to overcome. In the applied world, senior managers responsible for performance measurement often argue that randomization or expensive statistics are luxuries. We argue that this state of affairs is unnecessary; that is, it is possible to measure once and cut twice.

Impact thinking and tools

The most rigorous impact[ii] designs are anchored in counterfactual reasoning tools such as random assignment, linear mathematical models, and psychometrics – tools for examining differences between groups assumed implicitly to be homogenous. These models emphasize statistical power achieved through sufficient sample size, measurement using reflective items to maximize inter-item consistency, aggregation across cases (often ignoring nested structure) to produce estimates of average sample-level outcome effects, and analyses using one outcome variable at a time. These counterfactual reasoning tools for measuring impact, where impact is defined as the difference between two groups on one variable, are considered the gold standard in policy evaluation.[iii]

Finally, it is also worth noting that, in the education field, another nearly ubiquitous impact (non-experimental) design[iv] uses the tools of counterfactual reasoning to determine outcome effects for schools and, in most state accountability laws, distribute penalties to low performers. In short, many of the tools for measuring impact are not well-suited for the purposes to which they are put or the inherent complexities of the processes they are used to evaluate or understand.

CQI thinking and tools

CQI, like all work on organizational processes, requires hammer-nail reasoning; that is, identifying and measuring specific objects or processes hypothesized to have specific causal effects on other objects or processes. In CQI, data produced for improvement is most useful where (a) disaggregated to the lowest level of measurement (i.e., the item level) using formative measurement principles, (b) disaggregated for each case or *type* of case (e.g., each specific type of micro-setting and each specific type of participant), and (c) used to identify multivariate subgroups characterized by similar key operating characteristics (or, *components*) so that improvement responses can be efficiently targeted.

Hammer-nail reasoning[v] is anchored in the tools of developmental theory, qualitative and historical methods. However, the fields of psychology and education have over the years developed a science of categorical and ordinal description, pattern-centered and multi-level modeling, and formative measurement. These tools are designed to represent an object of measurement holistically for people and settings, to represent change in correspondence with actual individuals differing experiences, and to produces inferences about impact. The Quality-Outcomes design is a set of Hammer-nail reasoning tools.

Method

The primary purpose of the Quality-Outcomes (Q-O) evaluation design is to first differentiate intervention (e.g., summer learning offerings) subgroups by quality of instruction and then to compare types of individual student growth (e.g., pre-to-post change) across the quality subgroups. This "skill growth by levels of quality" design has been used with some frequency in early childhood evaluations (e.g., Karoly, 2014; Thornburg, Mayfield, Hawks, & Fuger, 2009) and was the subject of extensive study in the literature on aptitude-treatment interactions (Cronbach & Snow, 1977). This Q-O design does not align well to the tools of counterfactual reasoning that seek to equate groups at baseline using randomization and psychometrics. It aligns better with the tools of hammer-nail reasoning that seek to follow the sequence of causal events and processes and produce inferences through detailed description and replication.

Research questions

The methodology for the Quality-Outcomes design is guided by a sequence of approximately eight research questions. First, there are two questions about instructional quality that require description of the *baseline profile of service quality for each setting*; that is, the fidelity of implementation to the instructional model or standards in each setting:

1. What is the prevalence of low and high quality across settings?
2. Which indicators of high quality are missing across settings?

Second, there are two questions about student skill that require description of the *baseline profile of skill for each student*:

3. What is the distribution of low and high student skill across settings?
4. What is the distribution of student low baseline skill within each setting?

Third, there is a single question about impact – that is, *how do student skill sets change from baseline to time 2* – that requires description of scale reliability and subgroup assignment at both timepoints.

5. How do student skill sets change from baseline to time 2?

Fourth, there are two questions about *impact on skills*, defined as the difference in skill change for students who are exposed to high versus low quality settings. These questions represent a criterion validity test for the instructional model/standards, including for the setting's capacity to produce equity effects for students with greater SEL vulnerability.

6. Is exposure to high-quality instructional practices associated with greater skill change compared to exposure to low quality instructional practices?
7. Do students exposed to high-quality instructional practices, who were in the lower-skill subgroup at baseline, gain as much or more than students who were in the higher-skill subgroups?

Finally, there is a single question about the extent to which Q-O impacts (i.e., exposure to high quality and SEL skill growth) *predict improvements in academic performance* and other school-related outcomes. This question represents a stringent criterion validity test for the instructional model/standards:

8. Do students exposed to high-quality instructional practices and who experience substantial SEL skill growth demonstrate improvements in academic performance and other school-related outcomes in subsequent years?

Performance measures

In this section, we describe a set of performance measures for education-related settings that seek to use the Q-O design. Table 1 describes the characteristics of effective performance data in the Q-O design and shows the relative areas of focus in impact evaluation and performance measurement for CQI.

Table 1. Characteristics of Effective Performance Data

Quality-Outcomes Design	Impact Evaluation	*Reliable.* Data should be seen by all stakeholders as precise and factual due to standardization of measures/methods, clarity about the object and method of measurement, and repeated use of the instrument in field testing.
		Valid. Data are valid when they describe behaviors and conditions that are links in a causal chain of events desired by the actors involved (e.g., favoring insight about mechanism over prediction).
		Sensitive. Performance measures are focused on behaviors and conditions that are likely to change in response to interventions and can be used to describe change over a relevant performance period.
	Performance Measurement for CQI	*Timely*. Data that are available in real time as events occur, or just after completion, are more likely to hold relevance for actors.
		*Objective***.** Objective data are focused on behaviors and conditions that can be identified through observation and easily named in relation to practice.
		Feasible. Data collection must be feasible (i.e., the minimum data necessary are collected using typical organizational resources and from typical respondents).
		Multilevel. Ideally, measures should be designed to directly assess phenomena occurring at a specific level of the context or person. However, measures applied at lower-level units of analysis (e.g., staff) can be made useful at higher levels (e.g., organization) when aggregated across individual units to compose a higher-level rating. Rules for composition of information from lower-level units into representations of performance at higher levels require items that have an explicit theory for composition and rules for the necessary level of group agreement.
		Multipurpose. Performance data are multi-purpose when both data collection and data interpretation promote a shared language among actors and a framework to guide discussions about performance. In particular, observation-based data collection methods used by organizational staff build shared understanding of the objects of measurement and typical performance levels.

Four types of performance measures characterize the Q-O design: Quality of management practice, quality of instructional practice, student SEL growth, school outcomes.

Quality of management practice. Our recommended measures of manager practices include CQI implementation fidelity, workplace culture, and job satisfaction. Due to issues of response set and social desirability bias, we prefer staff reports about management practices as the source of the data; that is, rather than self-report we prefer an objective observer. A fidelity measure for the CQI process, which managers should be leading if performance information is to play a role in improving performance, is critical. The purposes of performance measurement cannot be served if the CQI cycle is not implemented.

Quality of instructional practice. Measures of teacher instructional practices are at the core of the Q-O design. In our prior work, this has been primarily with the Youth Program Quality Assessment (Smith & Hohmann, 2005), CLASS (Pianta & Hamre, 2009), ECERS (Harms and Clifford, 1980) CITE), or Preschool PQA (High/Scope, 2003), but several principles apply to the

numerous other teacher observation tools now on the market (e.g., Danielson, 1996, etc). First, the granularity of indicators really matters for both interpretability by users and for achievement of inter-rater reliability by observers. The appropriate level of granularity is situational, behavioral, usually social, and can be either momentary or more extended through time. If observers can reliably identify the behavioral situation, different aspects of behavior can be defined as indicators of high or low quality and identified reliably as present or absent in a given setting. However, this kind of indicator requires a methodological grounding in formative measurement (Diamantopoulos, Reifler, & Roth, 2008) and modeling with ordinal and categorical variables (Grice, 2015).

Student SEL skill growth. Student SEL skill growth requires measurement at two or more timepoints. Although there are many measures of SEL skill, we recommend behavioral indicators of mental skills (and teacher mental models of student mental skill) that reflect successful basic and advanced self-regulation of emotion, attention, and behavior. These elements – basic self-regulation, advanced self-regulation, and behavior – are the key components in integrative models that link aspects of setting quality to both performance and learning.

School outcomes. Academic achievement, SEL skills, grades, other school-related behavior (e.g., expulsions, suspensions), and subject-specific assessments are potential evidence of successful skill transfer, in this case from afterschool to school day settings. Establishing these outcome effects represents criterion validity in the Q-O design. Connecting the links in the hypothesized cascade of effects – exposure to high quality causes high SEL skill growth which transfers to school day outcome effects – creates opportunities to set validated benchmarks for performance.

Pattern centered analytics

Pattern-centered theory and methods are ideally suited for integrating the considerable amount of information reflected by the diverse range of measures included in these types of studies. For example, they reduce complexity without oversimplifying relations to only sample-level average effects and facilitate holistic representations by organizing multilevel multivariate data by reference to level-specific component processes (e.g., functionally interconnected variables) and their dynamic inter-level relations (e.g., cascading effects across levels), all of which tend to operate differently within different people and contexts, both within and across time.

We begin by identifying and measuring the key level-specific operating characteristics; that is, of the many possible measures of a complex system, such as an afterschool program, we focus on measures of the *components* of those system and avoid, as far as possible, measures of the *constituents* of those systems. Components are defined as part-processes expected to have causal force within the context of the system and questions being studied (e.g., active encouragement), and constituents are defined as part-processes unlikely to have causal force in relation to the goals of the study (e.g., the extent of hue in the green walls). This may seem like an obvious distinction, but making this distinction in practice may not always be so obvious, as explained by Bergman and Vargha (2013) in terms of the general measurement challenge:

> "An old Viking adage says, 'It takes a wise man (*sic*) to calculate fair shares of a loot'. Measurement is a fundamental activity ... and, like the wise Viking, it takes much deliberation to construct measurements that satisfy the different, sometimes conflicting, psychometric and conceptual demands. ...tailored to the specific research situation: there is no general recipe for how to do it" (Bergman & Vargha, 2013, pp. 13-14).

Assuming we have reliable and valid measures of the key level-specific components (e.g., several measures of staff practices and several measures of youth SEL skills), ideally at two or more points in time, we next focus on identifying the relatively-homogeneous subsets of persons and contexts described by the measures. This process involves applying a series of pattern-centered methods (e.g., hot deck imputation of missing data, identification and removal of extreme multivariate outliers, and cluster analysis) separately to each set of time-and level-specific variables.

The basic ideas here are (a) causal dynamics occur within specific places and times, (b), the relations among components tend to be more tightly coupled within specific levels and times than across levels and time, and (c) we can best understand the overall system dynamics by focusing first on holistic descriptions of components operating within specific levels (e.g., people or contexts) at specific points in time. An important corollary to these ideas is that, from a pattern-centered perspective, a score on a particular variable gets its meaning (for either persons or contexts) from its relation to scores on other component variables operating within the same person (or context), time, and level of analysis, not from other people's (or contexts') scores on that variable (cf. Magnusson, 2003).

The main goal of the initial set of pattern-centered methods is to identify a set of relatively-homogeneous subgroups of persons (or contexts) reflecting the dominant types of people (or contexts) included in the sample. Organizing multilevel, multivariate complexity by

reference to relatively-homogeneous subgroups allows to avoid the pitfalls associated with overly general sample-level averages and overly specific idiosyncratic particulars (cf. Kluckhohn and 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 the represent complex interactions characterizing proximally-integrated systems with fairly simple, pattern-centered, categorical variables.

Assuming we have used relevant component variables and identified a set of relatively-homogeneous subgroups, we are then in a position to examine a wide range of questions related to how these types of people and contexts (a) differ from each other, (b) interrelate with each other, and (c) move through time. The basic set of procedures involves various elaborations on Bergman et al.'s (2003) LICUR method (i.e., LInking ClUsters after removal of a Residue) and Cairns & Rodkin's (1998) *prodigal* method. The LICUR method can be used to identify both (a) higher-order, cross-level, configurations (e.g., specific kinds of people in specific kinds of contexts) and (b) individual- or context-level longitudinal pathways characterized by stability or a diverse range of possible changes. Using the results of the LICUR method applied to profiles or configurations from two or more points in time, the prodigal method shifts the focus from the full range of possible pathways (e.g., of SEL skill growth) to a series of focused contrasts between (a) individuals (or contexts) who follow the pathway that would be normatively expected given their initial profile or configuration pattern (e.g., low-skill youth remaining low-skill across time) and (b) individuals (or contexts) who deviate from the pathway that would be normatively expected given their initial profile or configuration pattern (e.g., low-skill youth who develop higher skills across time).

In addition to providing a powerful framework for identifying variables that predict or follow from such divergent (aka, off-diagonal) pathways (e.g., the extent to which such divergence is predicted by exposure to higher- or lower-quality instructional practice profiles), the prodigal method can be extended to provide substantial leverage for ruling out the possible influences of confounding variables (aka, endogeneity, or selection effects) by incorporating covariates and other predictors into, for example, logistic regression models applied to the initially homogeneous subgroup and their diverse pathways across time. Using this approach, we can examine the extent to which the off-diagonal pathway characterizing initially low-skill youth who develop higher SEL skills is predicted by any number of different profiles of instructional practices (or configurations of instructional practice profiles coupled with program support or

fidelity profiles), net the simultaneous influence of subtle differences in their initial SEL skill levels, instructional practices, program implementation fidelity, family SES, or any other substantive or demographic variable that represents a potentially alternative explanation for why some high-risk youth manage to defy normative expectations of stagnation where exposed to high-quality program or instructional practices.

Results

In this section, we use artificially produced data to visualization of that "made up" data to further illustrate the evidence produced from the Q-O design questions and methodology. These visualizations are designed to demonstrate the utility of the design and elaborate some aspects of methodology. Although these are not real data, we did try to reflect the patterns that we frequently see in real data which can be reviewed in many reports published at the David P. Weikart Center (e.g., Smith, Ramaswamy, Helegda, & Macleod, 2017; Smith, Roy, Peck, Macleod, & Helegda, 2017; Smith, Roy, Peck, & Macleod, 2018).

Useful information about practice

One of the key purposes of the Q-O design is to produce useful information for use by practioners during CQI cycles. A number of practice-relevant questions can be addressed by considering profiles of instructional practices. Figure 1 presents results from the pattern-centered sequence of analytics using measures for four hypothetical *domains of instructional quality*: Basic Safety (e.g., staff welcomes each student), Basic Learning Conditions (e.g., staff models skills), Social Interaction (e.g., teams pursue goals), and Advanced Learning Conditions (e.g., reflection). Each profile represents a distinct subgroup of programs with *distinct instructional approaches*: Low Quality, High Quality – Direct Instruction, and High Quality – Participatory Instruction. The *distribution of instructional approaches* across programs is also shown in Figure 1: 18% of summer settings were characterized by the lowest-performing profile, and 25% were characterized by the highest-performing profile. Sites characterized by the low-quality profile may not be producing positive effects on student learning and are obvious targets for resources and improvement. Sites characterized by the highest-quality profile represent exemplars of best practice fit to local circumstances and populations. It is also worthy of note that the shape of the profile is as important as summary scores. For example, we have learned from past experience with these kinds of data that the middle profile tends to be a "direct instruction" profile – with high learning supports but minimal supports for interaction and reflection/planning – that appears to be an intentional teaching style chosen by some staff.

Figure 1

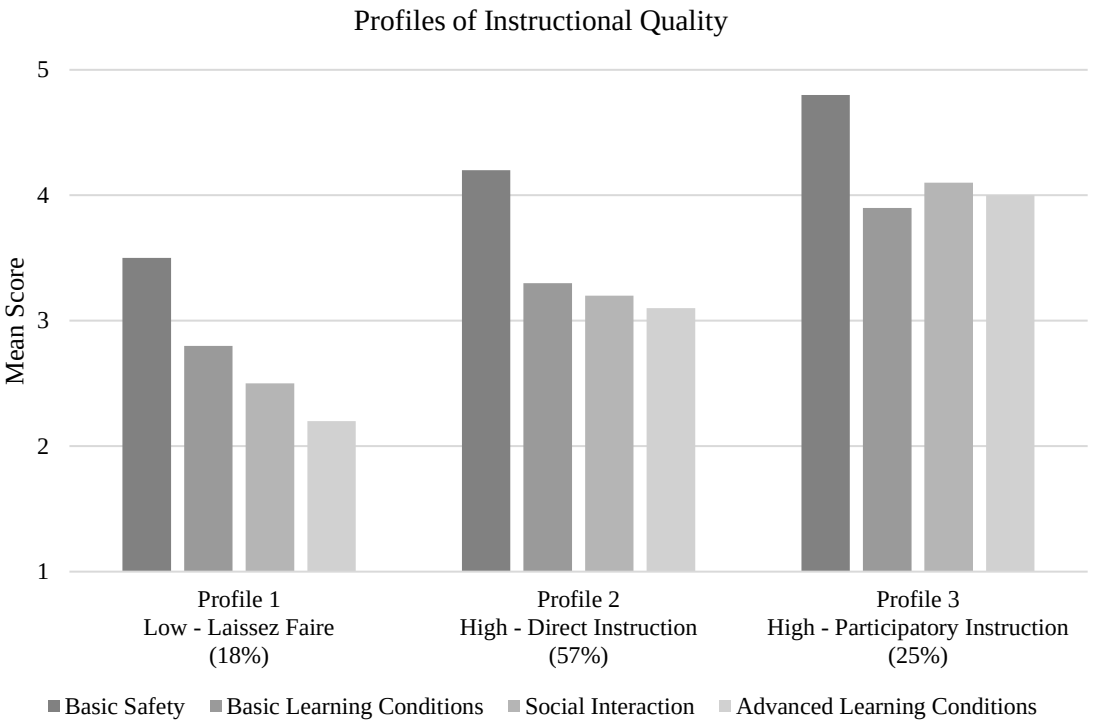


Table 2 shows another product of the Q-O design and performance measures for instructional process: low-scoring items from the observational measure of staff practices. Staff practices identified in the hypothetical data in Table 2 were not present in 30% or more of programs. These *infrequently used practices* are clear targets for improvement. In our experience, these kinds of measures are particularly helpful for understanding the level of adult responsiveness in settings trying to increase SEL supports.

Table 2. Low-scoring Items

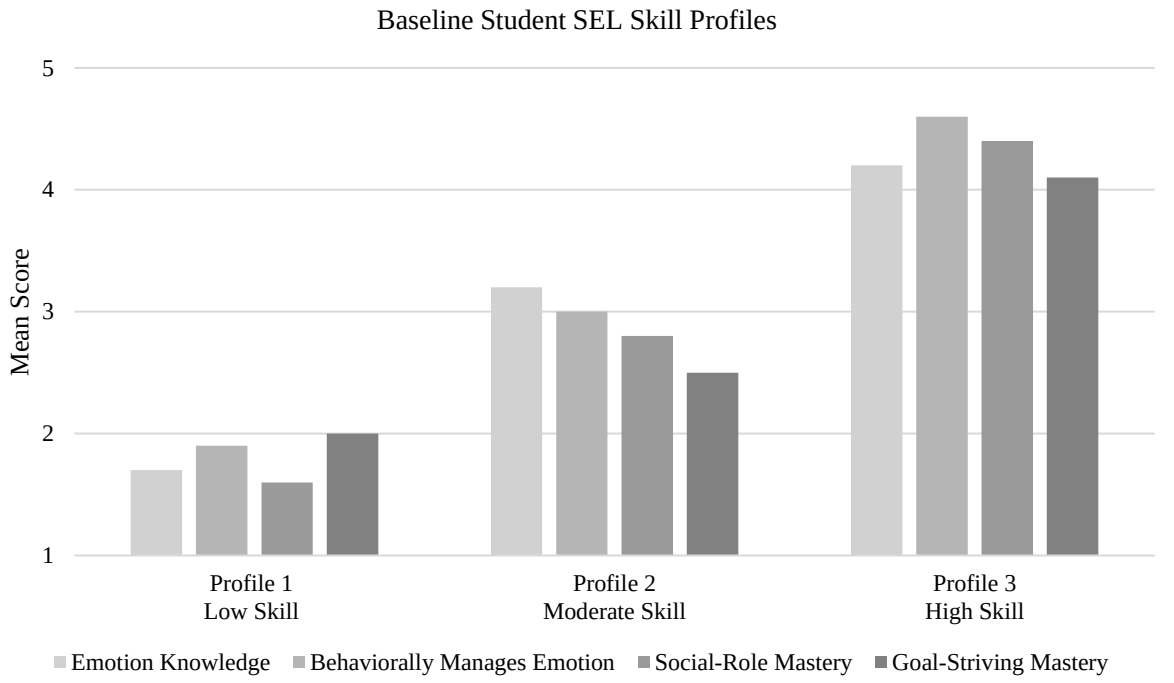
Item	Percent with score of 1
Staff provides a structured opportunity for youth to make plans (e.g. staff has youth write down their next steps for a project or have students converse about how they are going to accomplish their goal for the day.)	70%
Staff creates the opportunity of all youth to engage in a focused self-awareness exercise.	75%
Staff engages all youth in an intentional process of reflecting by practicing a specific reflection strategy.	68%
There is evidence of active inclusion and respect for all youth and there is no evidence of bias (based on culture, race, religion, ability, gender, sexual orientation, or ability) on the part of staff.	55%
All youth are individually greeted with genuine interest when entering the classroom (e.g. staff has some sort of greeting ritual or staff asks questions about youths' life,	62%

"How was your game this weekend?" "Is your sister feeling any better?")

Identify student skill building needs and the distribution of skills across settings

Figure 3 shows the results from pattern-centered analysis using individual student's scores for four teacher-rated SEL behavioral skills (Smith, 2013). Again, each profile represents a subgroup of students who have similar skills on each of *four SEL measures*: Emotion Knowledge (e.g., use emotion words), Behavioral Management of Emotion (e.g., manages hot emotions), Social Role Mastery (e.g., fulfills group roles), Goal-Striving Mastery (e.g., makes plans). From the *student SEL skill profiles* in Figure 3, it appears that students with very different skill levels were attending programs in the network, with about 53% of students (profile 1) struggling to self-regulate most of the time (i.e., mostly scoring under 3) during the program. In contrast, about 47% of students (profile 3) were successfully self-regulating most of the time (score over 3) during the program.

Figure 3



With this baseline knowledge of student SEL skill, it is also possible to see the *distribution of students in the lower SEL skill profile across sites*. In our experience, there is great variation in student SEL skills across settings. One implication of this variation is that, in a program setting where 70% of the students are not successfully self-regulating, that program setting has very

different and immediate (!) needs (e.g., more staff) compared to a setting where only 10% of students are having a difficult time engaging with the content.

Modeling impact and equity effects

The Q-O design starts with a simple *path impact model* for fitting data to tests of student outcome effects; in this case, the impact of exposure to high instructional quality on students SEL and school day outcomes. Figure 4 presents a heuristic example of how we evaluate relationships between staff instructional practice quality and student SEL skill growth. In the left-hand column of Figure 4, the active component of the setting (i.e., instructional quality) indicates high or low instructional quality. The top row indicates three different pathways of stability or change for each student. The interior cells represent the results of crossing different forms of instructional quality with different paths of SEL skill growth. The cell entries represent the proportion of students who evidence each form of SEL skill growth where exposed to each form of instructional quality. A chi-square test reveals the extent to which there are systematic relations between instructional quality and skill growth, and 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.

Figure 4. Path Impact Analysis

		Skill Growth		
		Positive	Stable	Negative
Quality	High			
	Low			

The logic of this path impact analysis can be extended by integrating profile and path variables for both setting and SEL skills into linear 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 profiles of instructional quality as opposed to 'pre-test' differences in potentially confounding variables.

The Q-O method also provides an approach to evaluating the equity of how outcome effects are distributed across students. We define equity as the extent to which settings support skill growth for students who entered the setting with low skills. Specifically, we are interested in the extent to which students with lower SEL skills, who also receive exposure to high quality

instructional practices, demonstrate greater skill growth compared to (a) lower skill students at baseline who were exposed to lower quality and (b) all other higher skill students at baseline who were exposed to high quality.

Working from the logic of the impact analysis illustrated in Figure 4, a *prodigal impact analysis* can be conducted by simply limiting the analysis to the subsample of students who were characterized by the lower-skill profile at baseline. Once each of these students has an identified pathway (i.e., positive, stable, negative), the proportion of students in each pathway can be crossed with any number of categorical variables, indicating how other aspects of management and instructional quality are related to the growth pathway of skill change.

Conclusions

The application of social science to social problems does not have to leave practitioners frustrated or doing guesswork about what the findings mean. We suggest that researchers can produce improvement value for clients, while advancing aspirations to greater certainty and evidence, using the Q-O design and its many possible variations.

This approach provides a feasibly path to both CQI uses and impact evaluation for outcome effects. However, in some cases it requires a step away from counterfactual reasoning and a step into hammer-nail reasoning, i.e., having a theory about how setting features and individual development interact at the proximal level to produce skill change – and then measuring a few of those key elements. Pattern centered analytics - crossing profiles for setting and skill gain information - make it feasible for a local organization to do both impact and CQI for their network and to replicate results over multiple program cycles. Reliability (e.g., consistency) then also becomes evidenced in the repeated pattern, i.e., moving from reliability of measures to reliability of model.

In our efforts to focus on methodology for performance measurement, we left the other major parts of the evidence based CQI model out, i.e., professional learning community and the logistics of the CQI cycle. Integrating expert practitioners into the performance measurement process dramatically increases the validity of the overall enterprise. For example, practitioner expertise is required to interpret and validate the profiles of program quality and student skill. The plans of expert practitioners for to change the situation requires their agreement with, and interpretation of (a) the meaning of the profiles and (b) the proportions of staff and youth in their organization who are characterized by those profiles.

Another critical point that was omitted from our discussion was implementation and cost. Although these issues have been addressed elsewhere (e.g., Grossman, Lind, Hayes, McKaken, &

Gersick, 2009; Smith, Peck, Macleod, Roy, Helegda, & Borah, 2018). It is worth noting that many school districts in the United States, in responding to their state ESSA plans, are currently collecting all of the performance data recommended as part of the Q-O design, making implementation of the design purely a matter of secondary data analyses.

References

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.

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.

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

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

Diamantopoulos, A., Reifler, P., & Roth, K. P. (2008). Advancing formative measurement models. *Journal of Business Research* (61:12), 1203-1218.

James-Burdumy, S., Dynarski, M., Moore, M., Deke, J., Mansfield, W., & Pistorino, C (2005). When schools stay open late: The national evaluation of the 21st century community learning centers program: Final report U.S. Department of Education, National Center for Education Evaluation and Regional Assistance. Available at http://www.ed.gov/ies/ncee.

Grice, J. W. (2015). From means and variances to persons and patterns. *Frontiers in Psychology, 6*.

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.

Harms, T., & Clifford, R. (1980). Early Childhood Environment Rating Scale (ECERS). New York: Teachers College Press.

High/Scope. (2003). *PQA preschool program quality assessment*. Ypsilanti, MI: High/Scope Educational Research Foundation.

Karoly, L. A. (2014). *Validation Studies for Early Learning and Care Quality Rating and Improvement Systems: A Review of the Literature*. https://www.rand.org/pubs/working_papers/WR1051.html

Klein, K. J., & Kozlowski, S. W. J. (2000). A multilevel approach to theory and research in organizations: Contextual, temporal, and emergent processes. In K. J. Klein & S. W. J. Kozlowski (Eds.), *Multilevel theory, research, and methods in organizations: Foundations, extensions, and new directions* (pp. 3-90). San Francisco, CA: Jossey-Bass.

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.

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.

McDavid, J. C., Huse, I., & Hawthorn, L. R. L. (2013). *Program Evaluation and Performance Measurement: An Introduction to Practice (2nd ed.)*. Thousand Oaks, CA: Sage Publications.

Pianta, R., & Hamre, B. (2009). Conceptualization, measurement, and improvement of classroom processes: Standardized observation can leverage capacity. *Educational Researcher, 38*, 109-119.

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.

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., & Hohmann, C. (2005). Full findings from the Youth PQA validation study. Ypsilanti, MI: *High/Scope Educational Research Foundation.*

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., 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., 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. Roy, L., Peck, S., Macleod, C. (2018). Evaluation of Program Quality and Social and Emotional Learning in American Youth Circus Organization Social Circus Programs. Washington DC: David P. Weikart Center for Youth Program Quality at the Forum for Youth Investment.

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., 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.

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.

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.

Appendix 1. Sample Performance Measures

Table reproduced from Smith, Roy, Peck, Macleod (2018).

Table 1. Performance Measures

System or Policy Level Practices
No system or policy-level data were collected.
Quality of Organization Practices
Vertical Communication: Manager provides feedback, is visible during the program, knows what is being accomplished, challenges staff, and makes sure program goals and priorities are clear. *Horizontal Communication:* Staff co-plan program policies or activities with other staff, discuss problems, and observe or are observed by other staff. *School-Day Content:* Staff are aware of school day academic content; coordinate afterschool activities with youths' homework; manage communication with parents, school day staff, and information sharing; and participate in meetings and parent-teacher conferences
Quality of Instructional Practices
Growth and Mastery Goals (Challenging Curriculum): Youth were exposed to new experiences, participated in successive sessions with increasing task complexity, were acknowledged for achievements, and identified what they are uniquely good at. *Instructional Total Score (Responsive Instruction):* Composed of ratings of staff instructional practice in three domains: A structured environment facilitated through guidance and encouragement (i.e., Supportive Environment), opportunities for leadership and collaboration (i.e., Interaction), and the capacity to promote planning and reflection (i.e., Engagement). *Youth Engagement:* Youth find activities important, use skills, have to concentrate, and experience moderate challenge.
Staff Rating of Youth Behavior (SRYB)
Emotion Knowledge: Youth identifies, names, and describes a wide variety of emotions. *Behaviorally Manages Emotion*: Youth behaves kindly, constructively, and non-defensively where confronted with both criticism and praise. *Social Role Mastery:* Youth helps others with tasks, roles, and responsibilities; seeks helps from staff when needed; and monitors team progress. *Goal Striving Mastery:* Youth evaluates options and potential solutions, creates plans, prioritizes tasks, manages time, and monitors goal progress.

Notes

[i] Also called "formative assessment" in the school day literature

[ii] Impact = two-group designs with experimental inference as the primary goal.

[iii] The intent to treat (ITT) variation on the counterfactual impact model carries this logic further, using assignment to treatment as the sole predictor variable used to define the division of the sample into comparable groups – regardless of the actual experience of participants. However, ITT impact estimates reflect mainly *decisions to invest* in programs; that is, they do not reflect the extent to which providers successfully *implement*, or youth *participate* in, these programs. Rather, they essentially assume complete implementation and participation, along with the inclusion of relevant measures. Given that (a) few programs are implemented at the highest level of quality, (b) substantial proportions of students do not attend these programs as much as they could, and (c) the "primary" variables tested are not necessarily things most likely to be influenced by the program, ITT impact estimates can be viewed as lower-bound estimates of the effects of the programs being tested. The number of problems and challenges associated with such ITT approaches are too numerous to list or discuss here but include issues such as (a) multifinality (e.g., different kids experience different things in response to the same instructional practices) and equifinality (e.g., different practices can be used to achieve the same effect), which are both pervasive phenomena that are masked and neglected where focusing on average effects that tend to apply to very few, if any, cases; (b) multivariate and multilevel program processes, which typically result in the omission of key mediator and moderator variables necessary for understanding and demonstrating how causal effects cascade (or dissipate) through complex systems; and (c) psychometric integrity, which requires reliable and valid items and scales that are aligned to specific objects within persons and contexts (e.g., well-validated psychological measures or behavioral indicators that closely reflect psychological processes), measurement models that are aligned to the objects of measurement (e.g., using a formative or configural measurement model instead of a purely reflective, composition model; Klein & Kowzlowski, 2000), and formative assessment and performance feedback procedures characterized by consequential validity.

[iv] In Campbell's terms, this is a one-shot, post-test only design, or a non-experimental design.

[v] Also referred to as physical cause of intentional behavior (Mohr, 1996) and Aristotle's final, formal and efficient causes (Rychlak, 1994).