(2016). Practical measurement and productive persistence: Strategies for using digital learning system data. *Journal of Learning Analytics*, *3*(2), 116–138. http://dx.doi.org/10.18608/jla.2016.32.6

Practical Measurement and Productive Persistence: Strategies for Using Digital Learning System Data to Drive Improvement

Andrew E. Krumm
SRI International, USA
andrew.krumm@sri.com

Rachel Beattie
Carnegie Foundation for the Advancement of Teaching, USA

Sola Takahashi
Carnegie Foundation for the Advancement of Teaching, USA

Cynthia D'Angelo
SRI International, USA

Mingyu Feng
SRI International, USA

Britte Cheng
SRI International, USA

ABSTRACT: This paper outlines the development of practical measures of productive persistence using digital learning system data. Practical measurement refers to data collection and analysis approaches originating from improvement science; productive persistence refers to the combination of academic and social mindsets as well as learning behaviours that are important drivers of student success within the Carnegie Foundation for the Advancement of Teaching's Community College Pathways (CCP) Network Improvement Community (NIC). Strategies for operationalizing non-cognitive factors using digital learning system data as well as approaches for utilizing them during improvement efforts are described.

Keywords: Improvement science, evidence-centered design, productive persistence.

1 INTRODUCTION

As interest in 21st century skills and non-cognitive factors increases, there is a growing need to measure these factors in valid, reliable, and actionable ways (Kosovich, Hulleman, Barron, & Getty, 2014). Researchers have developed self-report and observation instruments with high internal reliability and construct validity (Duckworth & Yeager, 2015), and while many instruments have attractive measurement properties, they are often time-intensive to complete and can require teachers and learners to disengage from learning activities in order to collect the desired information (Schraw, 2010).

Data collection techniques that keep teachers and learners engaged in the work of teaching and learning can be an important resource for improving schools (Yeager, Bryk, Muhich, Hausman, & Morales, 2013). One ready source of data that can be leveraged by both researchers and practitioners is generated from learners' direct interactions with digital learning tools and environments (Hadwin, Nesbit, Jamieson-Noel, Code, & Winne, 2007).

In this paper, we describe approaches for developing practical measures of hard-to-measure constructs, such as intrapersonal competencies and 21st century skills, using system log data (Stecher & Hamilton, 2014). Practical measures, also referred to as "improvement measures," are developed and used as part of quality improvement efforts and are distinguished from two other, perhaps more familiar measurement purposes: research and accountability. Data collected for the purposes of research and accountability are used to build knowledge and evaluate the performances of individuals and organizations, respectively. However, research and accountability measures can have limited meaning and utility to non-researchers for reasons that we detail below. As more and more attention is placed on developing students' 21st century skills and non-cognitive competencies, researchers and practitioners will need new approaches for collecting data in rapid and valid ways that can inform their efforts to develop these skills and competencies in learners.

As Stecher and Hamilton (2014) observe, a critical challenge in designing interventions to aid students in developing 21st century skills and competencies is measurement. Our focus in this paper is on the ways we as researchers developed measures that could aid instructors in community college developmental mathematics courses in delivering more targeted interventions to improve student persistence and use of good learning strategies, i.e., *productive persistence*. Productive persistence is related to what the National Research Council refers to as skills within the intrapersonal domain: "intellectual openness, work ethic and conscientiousness, and positive core self-evaluation. These clusters include competencies, such as flexibility, initiative, appreciation for diversity, and metacognition (the ability to reflect on one's own learning and make adjustments accordingly)" (2012, p. 4). *Persistence* addresses effort and continuing to work on a task, often in the face of initial failure (DiCerbo, 2015); *good strategies* represent behaviours, such as self-regulation and study strategies that can support the successful completion of a task (Zimmerman, 2002).

In this paper, we describe how we used online learning system data to develop behavioural measures of productive persistence within an overarching improvement effort directed at increasing the success rates of students enrolled in community college developmental math courses. Behavioural measures were developed as part of the Carnegie Foundation for the Advancement of Teaching's Community College Pathways (CCP) Network Improvement Community (NIC), which organizes colleges and universities throughout the United States around a common aim of tripling the success rate of learners in half the time of traditional developmental math offerings. Key to achieving this ambitious aim is implicating instructors, administrators, and researchers in the collective work of creating more effective learning environments using improvement science best practices (e.g., Langley et al., 2009).

Improvement science represents a different approach than traditional educational research, which typically focuses on what works (e.g., evaluation research) or why something works (e.g., theory building; Bryk, Gomez, Grunow, & LeMahieu, 2015). Importantly, improvement research uses both evaluation and theory-building to improve processes that occur within educational organizations (Russell, Jackson, Krumm, & Frank, 2013). However, improving processes in complex systems like schools requires not only knowledge of what works and why but also knowledge of how to integrate what works into regular organizational activities (Lewis, 2015; Penuel, Fishman, Cheng, & Sabelli, 2011). Based on the Institute for Healthcare Improvement's Model for Improvement, integrating what works can be focused on answering three questions: 1) What are we trying to accomplish? 2) How will we know that a change is an improvement? 3) What change can we make that will result in improvement? Answering these deceptively simple questions is complex work, and central to that work is the role of measurement.

2 PRACTICAL MEASUREMENT

Measures developed for the purposes of supporting improvement efforts involve different considerations and can take on different forms than those developed for the purposes of accountability or research (Solberg, Mosser, & McDonald, 1997). Over the last few decades, considerable effort has been invested in the development of accountability and research measures. In particular, the development of accountability measures has been prompted by the broader standards and accountability movement in the United States (Coburn & Turner, 2012). Accountability measures, such as graduation rates, student test scores, and teacher evaluation ratings, can be useful in defining the ultimate aims of improvement efforts, but they are often too general, infrequently collected, and distant from day-to-day activity to inform improvement work (Coburn & Turner, 2011). Research measures, which represent another common purpose for educational measurement, prioritize the disciplinary standards of a research community over practical use; moreover, they tend to be collected infrequently and are not necessarily designed to measure change over time (Solberg et al., 1997). The measures that most centrally benefit improvement efforts are "practical measures" — those taken directly from practice and easy to use in working to change educational processes (Bryk et al., 2015).

Practical measures are characterized by several qualities that distinguish them from measures for accountability or research (Yeager, Bryk, et al., 2013). First, practical measures are specific to an improvement effort. In contrast to accountability measures, which capture general system outcomes, practical measures provide information about the particular processes that educators aim to improve. Second, because practical measures are used to learn about specific changes to educational processes, they need to be sensitive to fluctuations in behaviours or performance. When changes to processes are tested, a practical measure should quantify the degree to which changes occurred. Third, in the role of providing feedback about specific changes, measures also need to be accessible in a timely manner to the practitioners implementing the changes.

Fourth, the effectiveness of practical measures depends on their usefulness as information to guide subsequent action (Bryk et al., 2015). Practitioners should be able to use these measures to inform decisions and next steps related to changes being tested. Moreover, practical measures should be meaningful to both practitioners and those in administrative roles supporting an improvement project. Therefore, practical measures depend upon the organizational routines enabling their use (Spillane, Parise, & Sherer, 2011). A culture of data use driven by judgment and evaluation is not conducive to the transparency necessary to use data in support of improvement. Instead, practical measures are most beneficial within a context where failures are seen as opportunities to learn, where trust pervades, and where practitioners are willing to take risks as well as make changes in their practice.

Fifth, practical measures should be based on a theory that outlines how changes to the system will lead to improvements in the service of an ultimate aim (Bennett & Provost, 2015). Without an improvement theory, practical measures are of limited value because the theory provides the rationale for why the measures are important to an overall improvement effort. By articulating a theory, aligning specific changes to that theory, and testing changes against measures organized around the theory, improvement teams can learn what is working in their local system, where there are breakdowns, and the conditions under which changes are not leading to improvements. Sixth, along with being based in a local improvement theory, effective practical measures are also predictive of the outcomes that practitioners are trying to improve. In this way, practical measures can serve as leading indicators of improvement goals (Provost & Murray, 2011).

Potential sources of practical measures in schools are data generated by student use of online learning systems (U.S. Department of Education, 2012, 2013). As more and more technology finds its way into schools under the banner of blended learning, there is a growing opportunity for both researchers and practitioners to take advantage of these data for research, accountability, and process improvement. For improvement purposes, online learning system data is attractive because it can be collected during learning activities without interrupting learners as well as analyzed rapidly to support formative decision-making (Nelson, Nugent, & Rupp, 2012). As we illustrate in the following sections, online learning system data can be used to develop behavioural measures that are predictive of long-term outcomes and useful for understanding changes made as part of an overall improvement effort.

3 PRODUCTIVE PERSISTENCE

The behavioural measures described in this paper were developed within an overall improvement effort directed at enhancing student outcomes in developmental math courses in community colleges throughout the United States. Upwards of 70% of students who enroll in a community college are assigned to developmental math; upwards of 80% of those students will not receive college-level credit even after three years (Bailey, Jeong, & Cho, 2010). Recognizing that the failure to serve developmental math students has numerous, interrelated causes, Carnegie developed Statway and Quantway to address multiple factors affecting student success. Statway is a two-term course that combines both developmental mathematics and statistical reasoning content. Quantway is a single-term, accelerated developmental math course paired with a college-level, credit-bearing quantitative reasoning course. Both Pathways emphasize the core mathematical skills needed for work, personal life, and citizenship. Moreover, both Pathways stress conceptual understanding and the ability to apply what is learned in a variety of contexts and problems. Cutting across both Pathways is a focus on rapid analytics, faculty development, network engagement, relevant content, and productive persistence, which combined, have contributed to tripling the success rate of students in less than half the time of traditional developmental math programs (Strother, Van Campen, & Grunow, 2013; Strother & Sowers, 2014; Van Campen, Sowers, & Strother, 2013).

Carnegie and the CCP NIC define productive persistence as *tenacity plus the use of good strategies*. Under this definition, productive persistence entails students having the academic and social mindsets as well as strategies and behaviours necessary to move effectively past challenges (see Figure 1). Academic and social mindsets, which are also referred to as intra-personal competencies (Stecher & Hamilton, 2014), are associated with student success in a variety of educational settings (Farrington et al., 2012) and can often be addressed by targeted interventions (Dweck, Walton, & Cohen, 2014). For example, it is common for a student in the U.S. to believe that he or she is not a “math person” when math does not come easily (Blackwell, Trzesniewski, & Dweck, 2007; Stigler & Heibert, 1999). Students may also question whether they truly belong or will be respected in a college setting — something that may especially beset students of colour, who may face negative stereotypes about the academic potential of members of their racial or ethnic group (CCCSE, 2010; Gardenshire-Crooks, Collado, Martin, & Castro, 2010; Walton & Cohen, 2007).

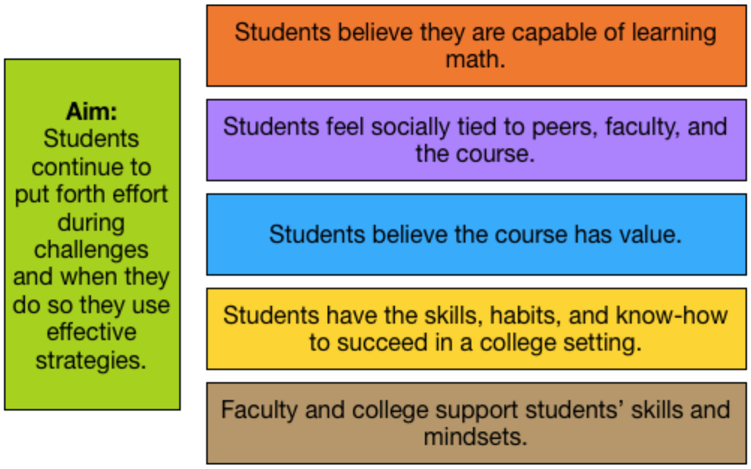
Aim: Students continue to put forth effort during challenges and when they do so they use effective strategies.
Students believe they are capable of learning math.
Students feel socially tied to peers, faculty, and the course.
Students believe the course has value.
Students have the skills, habits, and know-how to succeed in a college setting.
Faculty and college support students’ skills and mindsets.

Figure 1: Productive persistence driver diagram.

Based on multiple literature reviews as well as interviews with leading experts, faculty members, and students, researchers at Carnegie identified that many students who do not complete a developmental math course either withdraw effort or get too far behind during the first four weeks of a course (Cook, Purdie-Vaughns, Garcia, & Cohen, 2012; Vaquero & Cebrian, 2013). For that reason, Carnegie developed the Starting Strong Package to prepare faculty to launch a course successfully and help students develop the mindsets and skills needed to succeed in college. The Starting Strong Package is a combination of 10 instructor-led daily routines and activities launched during the first month of all CCP courses. The routines and activities aim to reduce stereotype threat and belonging uncertainty (Walton & Cohen, 2007), reduce mathematical anxiety (Jamieson, Mendes, Blackstock, & Schmader, 2010), increase the perceived relevance of mathematics and statistics (Hulleman & Harackeiwicz, 2009), and foster a sense of purpose (Yeager et al., 2014). It includes a brief, one-time "growth mindset" reading and writing activity that has been shown to increase overall mathematics grades among community college students (Yeager & Dweck, 2012; Yeager, Paunesku, Walton, & Dweck, 2013). There are also routines embedded into the online learning environment that promote the development of self-regulated learning strategies (Zimmerman, Moylan, Hudesman, White, & Flugman, 2011).

Measuring when students are not productively persisting is an increasingly important practice within the CCP NIC. Academic and social mindsets around productive persistence are currently measured in the CCP through a series of short, context sensitive self-report items given to students in the first and fourth weeks of the course (see Yeager, Bryk, et al., 2013, for a description of the development of these self-report measures). Behavioural measures stemming from online learning system data were intended to supplement pre-existing self-report measures within an overall measurement system used within the CCP NIC (Martin, Nelson, Lloyd, & Nolan, 2007).

When using online learning system data, there are few standardized processes for going from *clicks* to *constructs*. The relative newness of using online learning system data necessitated the identification of an approach that gives "communicable meaning" to behaviours operationalized using system log data (Provost & Murray, 2011). In our efforts to develop practical, behavioural measures of productive persistence, we used elements of the assessment process known as evidence-centred design (ECD) and the specific tool of design patterns to specify key elements of each measure.

4 EVIDENCE CENTRED DESIGN

Many digital learning environments generate large amounts of data on student interactions within the system, making it possible to track and identify an array of behaviours. To support our efforts in understanding student use of online systems within the CCP NIC, we utilized a series of ECD tools and processes. Mislevy and Haertel succinctly define ECD as follows: "Evidence-centred assessment design (ECD) provides language, concepts, and knowledge representations for designing and delivering educational assessments, all organized around the evidentiary argument that an assessment is meant to embody" (2006, p. v). Key to this definition is the idea of an "evidentiary argument," whereby assessment developers inherently make claims about what students know and can do based on the evidence collected from student performances. The aim of ECD is to promote the ideals of explicitness and reusability when making claims about students from often-imperfect evidence.

ECD is a process; it organizes the work of forming an evidentiary argument into five layers: 1) domain analysis, 2) domain modelling, 3) conceptual assessment framework, 4) assessment implementation, and 5) assessment delivery. *Domain analysis* involves understanding the focal domain of interest, such as scientific inquiry or self-regulated learning strategies. *Domain modelling* includes laying out one's argument for connecting claims about constructs to the tasks in which students are engaged and from which the evidence will be drawn. Design patterns are a specific tool used in the domain modelling layer; design patterns are built around three models and their interrelationships: 1) a student model (what knowledge, skills, and abilities are being measured?), 2) a task model (what activities will allow students to demonstrate those behaviours?), and 3) an evidence model (what data provide evidence — clicks, events, or log files — of those behaviours?). The *conceptual assessment framework* builds from the argument expressed in the domain-modelling layer as the actual algorithms and measurement models used to analyze collected evidence. *Assessment implementation* and *assessment delivery* involve preparing and deploying the assessment as well as collecting and analyzing data.

For the purposes of developing practical measures of productive persistence, we engaged in four ECD layers, paying particular attention to our evidentiary argument in the format of a design pattern. For the domain-modelling layer, SRI and Carnegie drew on existing research conducted by Carnegie on productive persistence; we also engaged in subsequent literature scans to better understand the types of behaviours identified using online learning system data. Following this step, we then aligned research-based behaviours and strategies in relation to the way in which learning tasks were organized within the online learning systems used in the CCP NIC. Lastly, we reconciled research-based behaviours and online learning system tasks with the actual data tracked and stored by the online systems. This reconciliation helped in framing our evidentiary argument for the analytical approaches (i.e., conceptual assessment framework) we used to distill each behaviour from the broader dataset (i.e., assessment implementation).

While ECD is often characterized as an approach for prospectively developing assessment tasks geared toward surfacing specific constructs, ECD and design patterns can also be used retrospectively to make sense of data generated by an online environment (DiCerbo, 2015). Using design patterns to reverse engineer pre-existing learning environments holds a great deal of potential for learning analytics researchers as design patterns draw attention to the ways in which certain digital learning environments could be used to measure specific constructs based on the data logged by the system and the structure of learning tasks within the system.

Given the new ways in which data can be collected and analyzed from digital learning environments, there is an increasing need to address the meaning behind these measures where millions of observations can be logged longitudinally. Measurement challenges, such as 1) under-representing a construct and 2) introducing construct-irrelevant variance (Messick, 1994), can come about regardless of the analytical technique used or the size of the dataset. Underrepresentation of a construct, for example, implies that the data used to measure a construct does not adequately address underlying components of the construct. This is particularly problematic when using online learning system data because, unless engineered around specific tasks to generate clear evidence (e.g., Cognitive Tutors), the learning environment may not collect nor store data on the specific behaviours a researcher might wish to use. Instead, researchers must make do with the data at hand, which could lead to using data unrelated to the construct. In providing a structured process, ECD, and in particular, design patterns, hold promise for developing measures using online learning system data because, as a process, ECD can help researchers see where and how specific behaviours can be measured within specific activities in digital environments (Mislevy, Behrens, DiCerbo, & Levy, 2012; Rupp et al., 2012).

5 BEHAVIOURAL MEASURES OF PRODUCTIVE PERSISTENCE

During the 2013–2014 academic year, Statway courses used Carnegie Mellon's Open Learning Initiative (OLI) platform. Using data from OLI as well as primary data collected by Carnegie researchers and community college faculty members, we distilled multiple practical measures associated with student use of the system. The types of events logged by OLI included pages viewed by students (PV), assessments referred to as "Checkpoints" (CP), and practice activities referred to as "Learning by Doing" (LBD) and "Did I Get This?" (DIGT). LBDs and DIGTs are interspersed throughout pages students are expected to read, and as their names imply, help students practice content-related knowledge and skills through interactive tasks (i.e., LBD) and self-assessments (i.e., DIGT). Statway I, which is often implemented in the first semester of an academic year, is typically comprised of six modules (e.g., "Types of Statistical Studies and Producing Data") and twelve topics (e.g., "Collecting Data — Sampling"). For the purposes of this paper, we report on behavioural measures developed using PV, CP, LBD, and DIGT events across six modules within Statway I.

Along with OLI data, Carnegie also collected common assessments from community colleges participating in the CCP NIC. The pre-assessment was developed to assess students' understanding of mathematical concepts (e.g., fractions, decimals, and algebraic notation) before entering a CCP course. The pre-assessment is comprised of 21 items with an internal reliability of .80. It is given to students in the first week of the course and has been identified as one of the best early predictors of student success in CCP courses. The summative assessment is made up of 40 items; it is also a rigorously validated and comprehensive instrument. For example, all items on the post-assessment were field-tested with over 500 members of a comparison sample who had taken a college mathematics or statistics course. A recent report by Strother and Sowers (2014) describes the relationship between higher summative assessments scores and students' end-of-term grades within the CCP. Both the pre-assessment and the summative assessment were scored using a 1-parameter item-response-theory model, i.e., a Rasch model.

5.1 Development Process

The dataset used to develop behavioural measures of productive persistence included the pre-assessment, the summative assessment, and OLI data, which included PV, CP, LBD, and DIGT events. This historical dataset provided the opportunity to identify predictive measures that could then be used by CCP instructors using real-time data displays in later semesters to measure ongoing improvement efforts. As we used the productive persistence framework to identify potential behaviours, make sense of OLI learning tasks, and analyze evidence generated from those tasks, we used the student, task, and evidence prompts from a typical ECD design pattern to organize the process. For example, an effective task model includes *characteristic features* and *variable features* associated with the task in which students are engaged along with specific *task products* (i.e., what students explicitly or implicitly produce as they engage in and complete a task). Characteristic features help in clarifying what is required of a task in order to identify a focal behaviour. For example, characteristic features of a task from which one might wish to measure persistence would require 1) a threshold for determining success, 2) student feedback in relation to that threshold, and 3) opportunities for students to retry the task. Variable features are aspects of the task that can vary or be varied intentionally. Variables such as difficulty of the task, whether a task is graded, and time allotted for the task can vary or be varied intentionally when measuring persistence.

The evidence model, unlike the task model, specifies how task products are analyzed in the form of *potential observations* and *potential frameworks*. Potential observations in the evidence model outline what to focus on across various task products along with specifications for the qualities of task products, such as high, medium, or low, as well as right or wrong. Lastly, *potential frameworks* help in putting qualitative assessments of evidence into context. For example, potential frameworks aid in articulating the conditions (e.g., early or late in a task) under which high or low is appropriate. Through this iterative and comparative process using design patterns, we began to identify potential behaviours that could be operationalized using OLI data elements in relation to the behaviours identified across multiple research scans.[1]

As we aligned student, task, and evidence models, we also engaged in preliminary descriptive and correlational analyses related to CPs, PVs, LBDs, and DIGTs. Early work, for example, pointed to the importance of completing and succeeding on CPs in relation to student performance on the summative assessment. With an initial understanding of the importance of passing CPs, we returned to the literature in an effort to better understand how to conceptualize that which students do before and after engaging in these assessments. Specifically, we used literature related to a “preparation for future learning” orientation, which addresses students’ abilities to adapt to new environments and act as independent learners (Bransford & Schwartz, 1999; Schwartz, Bransford, & Sears, 2005). We focused task and evidence models around the *choices* that students made prior to and after taking a CP. Schwartz and Arena (2013) argue that in learning environments where students have discretion, the activities and resources used in the service of learning can be more illustrative than traditional, item-based assessments in understanding students’ development as independent learners.

Schwartz, Arena, and colleagues prospectively design the environments that assess the choices that students make in order to evoke specific constructs. Thus, the choices that students make in terms of the activities and resources they use in completing a task are designed to elicit data that is interpretable in relation to the task and intended construct, such as critical thinking. Moreover, whether or not a

[1] Visit www.a4l.sri.com to view sample design patterns as well as the productive persistence design pattern.

learner is said to have demonstrated the construct combines system log data and an outcome measure, whereby the outcome, such as a win-state in a game, is not a standalone activity but embedded in the overall task. As Rupp et al. (2012) observe, the ability to see commonality across diverse tasks is one of the many benefits of using ECD. Within OLI, a CP is a standalone activity; reading and practice activities available to students are aligned with the CP and are at the student's discretion to use during an OLI session. These characteristic features of learning tasks within OLI provided the opportunity to make sense of the activities students engaged in prior to and after taking a CP.

The choices that students made prior to and following CPs were operationalized in the following ways. We first ordered the events that students generated across PV, LBD, DIGT, and CP activities by the timestamp logged by OLI within a session, defined as the time between a unique login and logout. We organized our overall task model around individual sessions, whereby students could exercise discretion in the activities in which they engaged. For example, students could take an assessment that had a threshold for success, retake assessments multiple times, and move freely among reading, practice, and assessment activities.

One of the first behaviours that we were interested in understanding was the degree to which students read and practiced during a session where they took one or more CPs. In some courses, CPs were required, so it was likely that we would see consistent CP use within these courses but that students might not engage in other types of activities. Thus, the first measure that we developed captured what we referred to as a *CP-only session*. CP-only sessions were defined as a student logging into OLI, taking one or more CPs, and not logging other activities during that same session. Preliminary analyses revealed that the more CP-only sessions that a student logged, the less well he or she did on the summative assessment.

The CP-only measure was one way of understanding the types of sessions that students logged. Related to CP-only sessions, we then developed a measure capturing whether students read, practiced, and assessed within a single session. We referred to these multi-activity sessions as *robust sessions*. For this measure, students needed to engage in PV, LBD/DIGT, and CP within the same session; however, we were agnostic to the order in which students engaged in these activities. We further operationalized student behaviours in relation to the number of events that a student logged before taking their first CP within a session. While robust sessions did not attend to the order of events, for this measure, referred to as *events before first CP*, we wanted to capture what students did in preparation for an assessment. Schwartz and Arena (2013) point to the importance of attending to what students do prior to an assessment as indicative of broader patterns of development in relation to becoming independent learners. The final behaviour that we illustrate in this paper captures how students end an OLI session. For this measure, referred to as *no attempt after low score*, we focused on when students scored below 60% on a CP and whether or not they engaged in any activity following a low score. Unlike CP-only sessions, robust sessions, and the number of events before taking a CP — all of which could be argued to be productive or less productive learning strategies, ending a session on a low score directly addressed the persistence element of productive persistence.

(2016). Practical measurement and productive persistence: Strategies for using digital learning system data. *Journal of Learning Analytics*, *3*(2), 116–138. http://dx.doi.org/10.18608/jla.2016.32.6

We assessed the predictive validity of the above behaviours in the follow forms: 1) the percentage of CP-only sessions logged across all sessions (PCPONLY), 2) the percentage of robust sessions, i.e., sessions in which students read, practiced, and took a CP, across all sessions (PROBUST), 3) average number of events, across all sessions, engaged in prior to taking the first CP within a session (AVGBFRCP), and 4) the percentage of times a student ended a session after scoring at or below 60% on a CP (PNOATT). Table 1 provides descriptive statistics for the two assessments, number of overall sessions logged by students, and behavioural measures. The four behaviour measures are reported as continuous variables in Table 1. For both modelling and interpretability reasons, we created categorical versions of these behavioural measures in our final predictive validity models.

Table 1: Descriptive statistics of assessments, number of sessions, and behaviours.

	Mean	SD	Min	Max
Pre-Assessment (PRE)	-.074	1.19	-3.02	3.78
Summative Assessment (SUM)	.72	.78	-1.32	3.65
Number of Sessions Logged (SESS)	19.14	14.55	0	160
% No Attempt After Low Score Sessions (PNOATT)	14.59	22.58	0	100
% Read, Practice, and Assess Sessions (PROBUST)	19.09	15.91	0	100
% CP Only Sessions (PCPONLY)	25.18	28.10	0	100
Average Number of Events Before First CP (AVGBFCP)	6.90	5.91	0	48

5.2 Assessing Predictive Validity of Behavioural Measures

The previous section described how we used ECD to align research-based behaviours and strategies to data generated by student use of OLI. After aligning student, task, and evidence models though design patterns, we identified whether behaviours predicted student performance in Statway I. Critical to helping instructors promote productive persistence is the degree to which behavioural measures attached to specific interventions are predictive, interpretable, and can support instructors in working with students in a more timely and targeted way.

To help in creating more interpretable measures, we created categorical versions of each variable that compared students who did not engage in the behaviour with students who did engage in the behaviour to some degree and those who did to a much higher degree. Categorical versions of variables also helped in overcoming more general interpretation issues in that the PNOATT, PROBUST, and PCPONLY behaviours had multiple students who did not evidence the behaviour. For PNOATT, 730 students who scored 60% or lower ended a session by engaging in another activity. However, 644 students ended a session, at least once, on that low score. Given the large proportion of students who did not evidence the behaviour, we created separate categories, where appropriate, for students who didn't engage in the behaviour in order to isolate the costs and benefits of engaging in these behaviours in the first place and to varying degrees.

For PNOATT, we created four categories. The first category, "No Low Score," was unique to this behaviour because it was premised on students experiencing a low score in the first place. Therefore, we treated students who did not score at or below 60% on a CP (n=92) as a separate group. Students who did score at or below 60% but did not end a session on a low score were coded as "None" (n=730). Those who did end a session on a low score were placed into two categories based on the percentage of sessions that they engaged in the behaviour. Students were place into two groups based on whether they were above or below the median for non-0% scores: "Below Median" and "Above Median." PROBUST and PCPONLY did not have similar issues around multiple interpretations for 0% scores; therefore, those who did not engage in the behaviour were coded as "None," and those who did engage in the focal behaviour were also placed into two groups based on whether they were above or below the median for non-0% scores. AVGBFCP did not have the same proportion of students who evidenced 0, so this behaviour was broken out into tertiles and coded as "Low," "Medium," or "High." Table 2 provides the degree to which students engaged in a focal behaviour for each re-coded category.

For students who experienced a low score on a CP, the median percentage of sessions was 25%, and students below this median (n=321) averaged 16.69% of their sessions whereby they ended on a low score. For the group of students above the median, they averaged more than three times the percentage of sessions per individual where they evidenced this behaviour (Mean%=50.57, n=323). Some 320 students did not have a robust session and were coded as "None." The median percentage of non-0% robust sessions was 23.26%, and students below this median averaged 14.60% robust sessions (n=578) and students above this median averaged 35.12% robust sessions (n=568). In a similar way, the percentage of students' overall sessions only comprised of CP (PCPONLY) events were as follows: "None" (n=358), "Below Median" (Mean%=11.33, n=560), and "Above Median" (Mean%=56.74, n=548). Lastly, the average number of events for the "Low" category of AVGBFCP was 1.38 (n=438), the "Medium" group (n=495) averaged 5.55, and the "High" group (n=488) averaged 14.03 events.

Table 2: Descriptive statistics for re-coded behavioural categories.

	Mean	SD	Min	Max	N
% No Attempt After Low Score Sessions (PNOATT)					
No Low Score	0	0	0	0	92
None	0	0	0	0	730
Below Median	16.69	5.36	6.25	25	321
Above Median	50.57	21.57	25	100	323
% Read, Practice, and Assess Sessions (PROBUST)					
None	0	0	0	0	320
Below Median	14.60	5.30	2.86	23.26	578
Above Median	35.12	11.72	23.26	100	568
% CP Only Sessions (PCPONLY)					
None	0	0	0	0	358
Below Median	11.33	6.98	0.81	26.32	560
Above Median	56.74	21.49	26.32	100	548
Average Number of Events Before First CP (AVGBFCP)					
Low	1.38	0.61	0	2.69	483
Medium	5.55	1.88	2.7	9.24	495
High	14.03	4.19	9.25	48	488

(2016). Practical measurement and productive persistence: Strategies for using digital learning system data. *Journal of Learning Analytics*, *3*(2), 116–138. http://dx.doi.org/10.18608/jla.2016.32.6

Our modelling strategy for assessing predictive validity involved using the above categorical measures of students' enactment of behaviours as predictors of their summative test performance (SUM). Along with behavioural measures, we included two other covariates: Z-scored number of sessions that the student logged during the semester (i.e., SESSZ) and student performance on a pre-assessment of mathematical conceptual knowledge (i.e., PRE). By modelling both sessions and students' incoming conceptual knowledge, we controlled for two factors that could contribute to student performance on the summative assessment and that were also valuable measures for instructors.

We first fit a null model with SUM as the dependent variable and no other covariates using a 2-Level Hierarchical Linear Model with students nested within instructors (HLM; Raudenbush & Bryk, 2002).[2] The null model assessed the degree to which student performance on SUM varied within and between instructors (Table 3). The model-based estimate for the grand mean of SUM (γ_{00}) was .74, which is nearly identical to the naïve average reported in Table 1. The between instructor variation around the grand mean (u_0) was .28. Therefore, instructors 2 standard deviations above mean or below the mean, on average, scored .56 logits higher or lower on SUM. At the student level (r), a student who scored 2 standard deviations above or below the mean scored 1.44 logits higher or lower. The intra-class correlation (ICC) for the null model (i.e., the degree to which students with the same instructor resembled one another) was 12.9%, signalling that most of the variation in SUM scores was attributable to student-level as opposed to instructor-level differences.[3]

Table 3: Null model.

Fixed Effect	Coefficient	Std. Err.	*t*-ratio	App. *d.f.*	*p*-value
SUM, γ_{00}	.741	.045	16.629	52	<0.001
Random Effect	SD	Variance	*d.f.*	χ^2	*p*-value
INTRCPT1, u_0	.280	.078	52	224.760	<0.001
level-1, r	.724	.524			

Note: 1,158 students and 53 instructors.

Table 4 presents the inclusion of baseline covariates, PRE and SESSZ. The following 2-Level HLM was fit:

$$SUM_{ij} = \gamma_{00} + \gamma_{10}{*}PRE_{ij} + \gamma_{20}{*}SESSZ_{ij} + u_{0j} + r_{ij}$$

SUM_{ij} is the summative assessment score for student i with instructor j. All predictors were grand mean centred; thus, γ_{00} is the intercept for students with an average pre-assessment (PRE) score and average number of Z-scored sessions (SESSZ). Both PRE and SESSZ were significant predictors. For example, a 1-unit change (i.e., logits) on PRE (γ_{10}) corresponded to a .307 unit change in SUM. Similarly, a 1-standard-deviation change in sessions logged on OLI (SESSZ, γ_{20}) corresponded to a .075 unit change in SUM.

[2] All models were fit using HLM 7.

[3] ICC = $u_0^2/(u_0^2+ r^2)$

Table 4: Baseline Covariates.

Fixed Effect	Coefficient	Std. Err.	*t*-ratio	App. *d.f.*	*p*-value
INTERCEPT, γ_{00}	.694	.039	17.894	52	<0.001
PRE, γ_{10}	.307	.019	16.447	985	<0.001
SESSZ, γ_{20}	.075	.026	2.887	985	0.004
Random Effect	**SD**	**Variance**	***d.f.***	χ^2	***p*-value**
INTRCPT1, u_0	.240	.058	52	193.965	<0.001
level-1, *r*	.646	.417			

Note: 1,040 students and 53 instructors; Robust standard errors reported.

The baseline covariates included in Table 4 were important for understanding the predictive validity of each behavioural measure, as we were interested in the degree to which each behaviour provided new insights into students' summative test performances. Including PRE and SESSZ, we fit 2-Level HLMs for each behavioural measure. Final models for each behavioural measure were specified in the following ways using mixed-effects notation:

Model 1: $SUM_{ij} = \gamma_{00} + \gamma_{10}*PRE_{ij} + \gamma_{20}*SESSZ_{ij} + \gamma_{30}*\boldsymbol{PNOATT0_{ij} + \gamma_{40}*PNOATT2_{ij} + \gamma_{50}*PNOATT3_{ij}} + u_{0j} + r_{ij}$

Model 2: $SUM_{ij} = \gamma_{00} + \gamma_{10}*PRE_{ij} + \gamma_{20}*SESSZ_{ij} + \gamma_{30}*\boldsymbol{PROBUST1_{ij} + \gamma_{40}*PROBUST2_{ij}} + u_{0j} + r_{ij}$

Model 3: $SUM_{ij} = \gamma_{00} + \gamma_{10}*PRE_{ij} + \gamma_{20}*SESSZ_{ij} + \gamma_{30}*\boldsymbol{PCPONLY1_{ij} + \gamma_{40}*PCPONLY2_{ij}} + u_{0j} + r_{ij}$

Model 4: $SUM_{ij} = \gamma_{00} + \gamma_{10}*PRE_{ij} + \gamma_{20}*SESSZ_{ij} + \gamma_{30}*\boldsymbol{AVGBFCP1_{ij} + \gamma_{40}*AVGBFCP2_{ij}} + u_{0j} + r_{ij}$

Table 5: Behavioural measures with baseline covariates.

Fixed Effect	Model 1	Model 2	Model 3	Model 4
INTERCEPT, γ_{00}	.704***	.682***	.697***	.689***
	(.038)	(.039)	(.040)	(.039)
PRE, γ_{10}	.286***	.303***	.304***	.305***
	(.018)	(.018)	(.019)	(.019)
SESSZ, γ_{20}	.087***	.071*	.038	.047
	(.027)	(.029)	(.029)	(.028)
PNOATT0 "No Low Score," γ_{30}	.405***			
	(.126)			
PNOATT2 "Below Med.," γ_{40}	-.187***			
	(.054)			
PNOATT3 "Above Med.," γ_{50}	-.142*			
	(.057)			
PROBUST1 "Below Med.," γ_{30}		.128^		
		(.070)		
PROBUST2 "Above Med.," γ_{40}		.295***		
		(.074)		
PCPONLY1 "Below Med.," γ_{30}			.046	
			(.053)	
PCPONLY2 "Above Med.," γ_{40}			-.143*	
			(.070)	

Fixed Effect	Model 1	Model 2	Model 3	Model 4
AVGBFCP2 "Medium," γ_{30}				.178** (.067)
AVGBFCP3 "High," γ_{40}				.215*** (.061)
Random Effect				
INTRCPT1, u_0	.231*** (.053)	.243*** (.059)	.249*** (.062)	.248*** (.061)
level-1, r	.636 (.405)	.639 (.408)	.642 (.412)	.641 (.411)

Note: ^<.1, *<.05, **<.01, ***<.001; For Fixed Effects ()=Standard Error; For Random Effects ()=Variance; Robust standard errors used for significance tests.

Based on the above model specifications, γ_{00} represents the mean of SUM_{ij} across the grand mean for all covariates. The variance of γ_{00}, $u_{0j,}$ is therefore interpreted as variance between instructors across the adjusted means. Each γ_{n0} represents the adjusted mean for a given covariate, and covariates, such as $SESSZ_{ij}$, represent the score for student *i* with instructor *j*, and r_{ij} represents the student-level residual variance. As noted above, PNOATT was modelled using four as opposed to three distinct categories given the multiple interpretations for 0% values. All other behaviours were modelled as three categories. For all models, "None" (i.e., PNOATT, PROBUST, and PCPONLY) or "Low" (i.e., AVGBFCP) served as the reference category. As with previous models, SUM_{ij} represented the summative assessment score for each student *i* with instructor *j*, and all covariates, including categorical variables were grand mean centred. Therefore, γ_{00} across all models represented the average SUM score for students with average PRE and SESSZ values, irrespective of behavioural category. We chose this centring strategy in order to compare intercepts across models. If we chose not to centre categorical variables, then the intercept would have represented the average SUM score, dependent upon other covariates, for students in the reference category, which was "None" or "Low" depending upon the behaviour, which would have led to multiple values for the intercept across models including the null model (Table 2) and baseline covariate model (Table 3). We explored both group- and grand-mean centring strategies, and the Deviance Likelihood Ratio Test indicated that grand-mean centred versions of the model fit the data better than group-mean centred models.

Table 5 presents the final models for each behavioural measure. Across all models, PRE was a significant predictor with similar orders of magnitude. The number of sessions that a student logged (SESSZ), however, was only significant for Models 1 and 2. For Model 1, all three categories of PNOATT were significant. Students who did not have a low CP score ("No Low Score," γ_{30}=.405***), scored significantly higher on SUM than students who did not end a sessions on a low score, i.e., the reference category. Those who did end sessions on a low score, on average, did less well on SUM (PNOATT "Below Med.," γ_{40}=-.187***; PNOATT "Above Med.," γ_{50}=-.142*).

For Model 2, students whose percentage of robust sessions was above the median, on average, performed higher on SUM (PROBUST “Above Med.,” γ_{40}=.295***). Aside from students not experiencing a low CP score (PNOATT “No Low Score,” γ_{30}=.405***), students experiencing a high proportion of robust sessions had the largest effect on SUM, controlling for all other factors. For Model 3, the percentage of CP-only sessions (PCPONLY) was significant for the above median category (PCPONLY2 “Above Med.,” γ_{40}=-.143*). For the same model but with the highest group, “Above Med.,” held out as the reference category, “None” and “Below Med.” were also found to be different (“None,” γ_{30} =.143, p=.041; “Below Med.,” γ_{40}=.189, p=.004). Lastly, an increase in the average number of activities a student logged before taking a CP across all sessions was related to increased SUM scores, with students in both groups scoring higher on SUM (AVGBFCP2 “Medium,” γ_{30}=.178**; AVGBFCP3 “High,” γ_{40}=.215***).

Across the four models described above, the highest categories (i.e., “Above Med.” or “High”) were significantly related to SUM, controlling for PRE and SESSZ. For PROBUST and PCPONLY, the below median categories, were not significant at .05 level; however, the directionality of the parameter estimates correspond to the general observation that across all behaviours the more students engaged in the behaviour, the stronger the effect on summative test performance. The statistical models presented above represent one element of validity, namely predictive validity, and the preceding discussion on how the behavioural measures were constructed using ECD and design patterns speak to the construct validity. Each aspect of the work adds to the evidentiary argument around measuring behavioural elements of productive persistence.

To summarize, the above analyses addressed both productive and persistence behaviours based on data collected from student use of an online learning system across an entire semester. For productive behaviours, we developed a measure that captured the degree to which students availed themselves of multiple types of activities within a given session. Another productive behaviour entailed reading and practicing before taking an assessment for the first time. Both behaviours, when engaged in across multiple sessions, were related to higher performances on the summative assessment. While these two behaviours are potentially productive, we also developed measures of potentially less productive behaviours, such as logging into the online system and only taking assessments, whereby students who consistently engaged in this behaviour did less well than students who did not. To capture aspects of student persistence, we measured the degree to which students ended a session on a score of 60% or lower without engaging in any further activity. Students who were consistently less persistent, did less well, on average, on the summative assessment than students who demonstrated persistence on a more consistent basis.

For practical measurement purposes, each behavioural measure, described above, was predictive of distal outcomes and each captured easily understandable, composite behaviours that could be tracked over time as well as used to measure formative processes associated with an improvement effort. Importantly, each measure was gathered from data tracking actual student behaviours without interrupting their learning.

6 DISCUSSION AND CONCLUSION

Supporting students' development of 21st century skills requires intentional effort (NRC, 2012). While commonsensical, the ways in which efforts to improve specific skills are organized (e.g., evaluation, research, or improvement) plays a significant role in how measurement is positioned in that work (Duckworth & Yeager, 2015). In this paper, we described the ways in which data from digital learning environments can be used to develop practical measures of *productive persistence*, defined by the Carnegie Foundation for the Advancement of Teaching as *tenacity plus the use of good strategies*. Productive persistence is a key driver within Carnegie's Community College Pathways (CCP) Network Improvement Community (NIC). The way in which we developed behavioural measures has potential implications for the field of learning analytics in supporting improvement science research. For example, learning analytics researchers potentially have a unique role to play in leveraging previously untapped sources of data from learning technologies, longitudinal data systems, and student information systems to both better understand educational processes and provide just-in-time measures for practitioners.

Two approaches informed the development work described in this paper: improvement science and evidence-centred design (ECD). From an improvement science perspective, we developed measures linked to an evidence- and practice-based theory for improvement (see Figure 1), tracked key processes, and were predictively valid. We used the principles of ECD and design patterns to align potential behaviours and strategies with tasks and evidence stemming from student engagement with an online learning system. The overarching aim for these measures was to help instructors participating in the CCP NIC track student behaviours, and we are currently engaged in co-development activities around instructor-facing data visualizations and interventions tied to co-developed visualizations.

Throughout this paper, predictive validity was held out as an important component of practical measurement because predictive measures of valued outcomes can increase practitioner motivation to use data in support of their improvement work (Provost & Murray, 2011; Solberg et al., 1997). Along with predictive validity, theoretically grounded measures can also support practitioner perceptions of value in using a measure over time. For example, based on the success of self-report productive persistence measures used in CCP NIC courses (see Yeager, Bryk, et al., 2013), predictive validity and a robust research base — combined — promoted initial as well as sustained use of the measures by instructors and researchers.

While predictive validity was established using a summative test of mathematical knowledge, it was not the only outcome explored, nor the only outcome valued by researchers or instructors. In order to be valuable for improvement purposes, each measure had to predict outcomes associated with the CCP NIC's overall improvement goals of increasing the number of students who complete developmental math courses. Therefore, we explored outcomes such as grades and course completion. Establishing the link between behavioural measures and an outcome is crucial in that there are many opportunities to measure student use of the online system and only one opportunity to measure how well students did in a course. By developing predictive measures of distal outcomes, change ideas attempted by instructors as part of the overall NIC can be assessed using these measures while there is still time to learn and improve.

To develop practical measures of productive persistence, we leveraged a historical dataset to engineer features in relation to known outcomes. Along with known outcomes, the historical dataset included a pre-assessment of students' incoming conceptual mathematical knowledge. During early descriptive and statistical analyses, we observed relationships between student scores on the pre-assessment and the number of reading and practice activities in which they engaged, whereby students with higher incoming content knowledge engaged with the system to higher degrees and in different ways. The relationships between domain knowledge and use of the system not only highlights the utility of including the pre-assessment in our predictive models (see Tables 3 and 4), it also points to potentially important future research directions on the relationships between content knowledge and 21st century skills.

There are important limitations to the measures described in this paper. First, the measures developed used historical data. As historical data, important factors affecting both student use of the online system as well as their overall performance in a course could be missing. Second, the four behavioural measures are specific to the improvement work of the CCP NIC and are not intended to support theory building or generalize beyond the CCP NIC. Future work, however, is directed at surfacing new behavioural measures and validating the ones outlined in this paper using data from new students. While these measures, in their current form, are limited in their generalizability, they can serve as example behaviours for other improvement and research teams to use, and the process by which the current measures were developed and validated could also serve as a ready approach for future teams.

Another important limitation of the measures described in this paper is that, even as a whole, they only measure certain elements of productive persistence, namely, student use of the online learning system. Productive persistence entails multiple beliefs about oneself as a learner and as a member of a learning community along with how one engages in learning activities (see Yeager, Bryk, et al., 2013). Multiple learning behaviours were not measured, as they were not captured by the online learning system. Therefore, the data reported in this paper represent potential elements of an overall measurement system that can be used to develop and test strategies that instructors could enact in their courses to promote students' productive persistence.

As noted above, the types of behaviours identified and then validated were based on data already collected and stored by the online system. We used evidence-centred design (ECD) as an overarching approach along with relevant prior literature in making sense of how students could use the online learning system in relation to evidence collected by the system. In this way, ECD was used to reverse-engineer potential behaviours from what was available. However, ECD is more commonly used as a prospective approach, whereby designers engineer specific tasks to generate specific evidence in support of making claims related to specific constructs.

One benefit of practical measures stemming from student use of an online learning system is that data can be collected frequently and without interrupting teaching and learning activities. While there are benefits to using online learning system data, practical measures are not limited to certain data collection approaches; they are recognizable in terms of their utility to support frontline workers learn and improve. A fundamental tenet of improvement science is that variation in outcomes is the key problem to solve (Bryk et al., 2015; Langley et al., 2009). Data from currently untapped sources hold great potential for both the fields of improvement science and learning analytics in measuring and positively intervening in these sources of variation.

ACKNOWLEDGEMENTS

This paper is based upon work supported in part by the National Science Foundation (NSF) (SMA-1338487) and SRI International. Any findings or opinions expressed in this paper are those of the authors and do not necessarily reflect the views of NSF or SRI.

REFERENCES

Bailey, T., Jeong, D. W., & Cho, S. (2010). Referral, enrollment, and completion in developmental education sequences in community colleges. *Economics of Education Review, 29*(2), 255–270. http://dx.doi.org/ 10.1016/j.econedurev.2009.09.002

Bennett, B., & Provost, L. (2015, July). What's your theory? Driver diagram serves as tool for building and testing theories for improvement. *Quality Progress*, 36–43.

Blackwell, L., Trzesniewski, K., & Dweck, C. S. (2007). Implicit theories of intelligence predict achievement across an adolescent transition: A longitudinal study and an intervention. *Child Development*, *78*, 246–263. http://dx.doi.org/ 10.1111/j.1467-8624.2007.00995.x

Bransford, J. D., & Schwartz, D. L. (1999). Rethinking transfer: A simple proposal with multiple implications. *Review of Research in Education, 24*(1), 61–100.

Bryk, A. S., Gomez, L. M., Grunow, A., & LeMahieu, P. G. (2015). Learning to improve: How America's schools can get better at getting better. Cambridge, MA: Harvard Education Press.

CCCSE (Center for Community College Student Engagement). (2010). *The heart of student success: Teaching, learning, and college completion (2010 CCCSE Findings)*. Austin, TX: The University of Texas at Austin, Community College Leadership Program.

Coburn, C. E., & Turner, E. O. (2011). Research on data use: A framework and analysis. *Measurement, 9*, 173–206. http://dx.doi.org/10.1080/15366367.2011.626729

Coburn, C. E., & Turner, E. O. (2012). The practice of data use: An introduction. *American Journal of Education, 118*(2), 99–111. http://dx.doi.org/10.1086/663272

Cook, J. E., Purdie-Vaughns, V., Garcia, J., & Cohen, G. L. (2012). Chronic threat and contingent belonging: Protective benefits of values affirmation on identity development. *Journal of Personality and Social Psychology*, *102*, 479–496. http://dx.doi.org/10.1037/a0026312

Duckworth, A. L., & Yeager, D. S. (2015). Measurement matters: Assessing personal qualities other than cognitive ability for educational purposes. *Educational Researcher, 44*(4), 237–251. http://dx.doi.org/10.3102/0013189X15584327

DiCerbo, K. E. (2015). Assessment of task persistence. In Y. Rosen, S. Ferrara, & M. Mosharraf (Eds.) *Handbook of research on computational tools for real-world skill development* (pp. 780–806). Hershey, PA: IGI Global.

Dweck, C., Walton, G., & Cohen, G. (2014). *Academic Tenacity: Mindsets and Skills that Promote Long-Term Learning* (Research Report). Seattle, WA: Bill & Melinda Gates Foundation.

Farrington, C. A., Roderick, M., Allensworth, E., Nagaoka, J., Keyes, T. S., Johnson, D. W., & Beechum, N. O. (2012). *Teaching adolescents to become learners. The role of non-cognitive factors in shaping school performance: A critical literature review.* Chicago, IL: University of Chicago Consortium on Chicago School Research.

Gardenshire-Crooks, A., Collado, H., Martin, K., & Castro, A. (2010). *Terms of engagement: Men of color discuss their experiences in community college*. New York/Oakland, CA: MDRC.

Hadwin, A. F., Nesbit, J. C., Jamieson-Noel, D., Code, J., & Winne, P. H. (2007). Examining trace data to explore self-regulated learning. *Metacognition and Learning*, *2*(2–3), 107–124. http://dx.doi.org/10.1007/s11409-007-9016-7

Hulleman, C. S., & Harackiewicz, J. M. (2009). Promoting interest and performance in high school science classes. *Science*, *326*, 1410–1412. http://dx.doi.org/10.1126/science.1177067

Jamieson, J. P., Mendes, W. B., Blackstock, E., & Schmader, T. (2010). Turning the knots in your stomach into bows: Reappraising arousal improves performance on the GRE. *Journal of Experimental Social Psychology, 46*, 208–212. http://dx.doi.org/10.1016/j.jesp.2009.08.015

Kosovich, J. J., Hulleman, C. S., Barron, K. E., & Getty, S. (2014). A practical measure of student motivation: Establishing validity evidence for the expectancy-value-cost scale in middle school. *The Journal of Early Adolescence, 35*(5–6), 790–816. http://dx.doi.org/10.1177/0272431614556890

Langley, G. J., Moen, R. D., Nolan, K. M., Nolan, T. W., Norman, C. L., & Provost, L. P. (2009). *The improvement guide*. San Francisco, CA: Jossey-Bass.

Lewis, C. (2015). What is improvement science? Do we need it in education? *Educational Researcher*, *44*(1), 54–61. http://dx.doi.org/10.3102/0013189X15570388

Martin, L. A., Nelson, E. C., Lloyd, R. C., & Nolan, T. W. (2007). *Whole System Measures* (IHI Innovation Series white paper). Cambridge, MA: Institute for Healthcare Improvement.

Messick, S. (1994). The interplay of evidence and consequences in the validation of performance Assessments. *Educational Researcher, 23*(2), 13–23. http://dx.doi.org/10.3102/0013189X023002013

Mislevy, R., & Haertel, G. (2006). *Implications of evidence-centered design for educational testing* (Draft PADI Technical Report 17). Menlo Park, CA: SRI International.

Mislevy, R. J., Behrens, J. T., DiCerbo, K. E., & Levy, R. (2012). Design and discover in educational assessment: Evidence-centered design, psychometrics, and educational data mining. *Journal of*

Educational Data Mining, *4*(1), 11–48. Retrieved from http://educationaldatamining.org/JEDM13/index.php/JEDM/article/view/22

NRC (National Research Council). (2012). J. W. Pellegrino & M. L. Hilton (Eds.). *Education for life and work: Developing transferable knowledge and skills in the 21st century*. Washington, DC: Committee on Defining Deeper Learning and 21st Century Skills, Board on Testing and Assessment and Board on Science Education, Division of Behavioral and Social Sciences and Education, The National Academies Press.

Nelson, B., Nugent, R., & Rupp, A. A. (2012). On instructional utility, statistical methodology, and the added value of ECD: Lessons learned from the special issue. *Journal of Educational Data Mining,* *4*(1), 224–230. Retrieved from http://www.educationaldatamining.org/JEDM/index.php/JEDM/article/view/27

Penuel, W. R., Fishman, B. J., Cheng, B. H., & Sabelli, N. (2011). Organizing research and development at the intersection of learning, implementation, and design. *Educational Researcher, 40*, 331–337. http://dx.doi.org/10.3102/0013189X11421826

Provost, L. P., & Murray, S. K. (2011). *The health care data guide: Learning from data for improvement*. San Francisco, CA: Jossey-Bass.

Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models in social and behavioral research: Applications and data analysis methods* (2nd ed.). Thousand Oaks, CA: Sage Publications.

Rupp, A. A., Levy, R., DiCerbo, K. E., Sweet, S. J., Crawford, A. V., Calico, T., Benson, M., Fay, D., Kunze, K. L., Mislevy, R. J., & Behrens, J. T. (2012). Putting ECD into practice: The interplay of theory and data in evidence models within a digital learning environment. *Journal of Educational Data Mining,* *4*(1), 49–110. Retrieved from http://www.educationaldatamining.org/JEDM/index.php/JEDM/article/view/23

Russell, J., Jackson, K., Krumm, A., & Frank, K. (2013). Theories and research methodologies for design-based implementation research: Examples from four cases. In B. J. Fishman, W. R. Penuel, A. R. Allen, & B. H. Cheng (Eds.), Design based implementation research: Theories, methods, and exemplars. *National Society for the Study of Education Yearbook*, *112*(2), 157–191. New York: Teachers College Press.

Schwartz, D. L., & Arena, D. (2013). *Measuring what matters most: Choice-based assessments for the digital age*. Cambridge MA: MIT Press.

Schwartz, D. L., Bransford, J. D., & Sears, D. (2005). Efficiency and innovation in transfer. In J. P. Mestre (Ed.), *Transfer of learning from a modern multidisciplinary perspective* (pp. 1–51). Greenwich, CT: Information Age.

Schraw, G. (2010). Measuring self-regulation in computer-based environments. *Educational Psychologist, 45*(4), 258–266. http://dx.doi.org/10.1080/00461520.2010.515936

Solberg, L., Mosser, G., & McDonald, S. (1997). The three faces of performance measurement: Improvement, accountability, and research. *The Joint Commission Journal on Quality Improvement*, *23*(3), 135–147.

Spillane, J. P., Parise, L. M., & Sherer, J. Z. (2011). Organizational routines as coupling mechanisms: Policy, school administration, and the technical core. *American Educational Research Journal, 48*(3), 586–619. http://dx.doi.org/10.3102/0002831210385102

Stecher, B. M., & Hamilton, L. S. (2014). *Measuring hard-to-measure student competencies: A research and development plan*. Santa Monica, CA: RAND Corporation.

Stigler, J. W., & Heibert, J. (1999). *The Teaching Gap*. New York: Free Press.

Strother, S., & Sowers, N. (2014). *Community college pathways: Summative assessments and student learning.* Stanford, CA: Carnegie Foundation for the Advancement of Teaching. Retrieved from http://www.carnegiefoundation.org/resources/publications/community-college-pathways-summative-assessments-student-learning/

Strother, S., Van Campen, J., & Grunow, A. (2013). *Community college pathways: 2011–2012 descriptive report*. Stanford, CA: Carnegie Foundation for the Advancement of Teaching. Retrieved from http://www.carnegiefoundation.org/sites/default/files/CCP_Descriptive_Report_Year_1.pdf

U.S. Department of Education. (2012). *Enhancing teaching and learning through educational data mining and learning analytics*. Washington, DC: Author. Retrieved from http://www.ed.gov/edblogs/technology/files/ 2012/03/edm-la-brief.pdf

U.S. Department of Education. (2013). *Expanding evidence approaches for learning in a digital world*. Washington, DC: Author. Retrieved from http://www.ed.gov/edblogs/technology/files/2013/02/Expanding-Evidence-Approaches.pdf

Van Campen, J., Sowers, N., & Strother, S. (2013). *Community college pathways: 2012–2013 descriptive report.* Stanford, CA: Carnegie Foundation for the Advancement of Teaching. Retrieved from http://www.carnegiefoundation.org/resources/publications/community-college-pathways-2012-2013-descriptive-report/

Vaquero, L. M., & Cebrian, M. (2013). The “rich club” phenomenon in the classroom. *Nature (Scientific Reports), 3*, 1174. http://dx.doi.org/10.1038/srep01174

Walton, G. M., & Cohen, G. L. (2007). A question of belonging: Race, social fit, and achievement. *Journal of Personality and Social Psychology, 92*, 82–96. http://dx.doi.org/10.1037/0022-3514.92.1.82

Yeager, D. S., & Dweck, C. S. (2012). Mindsets that promote resilience: When students believe that personal characteristics can be developed. *Educational Psychologist, 47,* 1–13. http://dx.doi.org/10.1080/00461520.2012.722805

Yeager, D. S., Paunesku, D., Walton, G., & Dweck, C. S. (2013). How can we instill productive mindsets at scale? *A review of the evidence and an initial R&D agenda*. A White Paper prepared for the White House meeting on “Excellence in Education: The Importance of Academic Mindsets.”

Yeager, D. S., Bryk, A., Muhich, J., Hausman, H., & Morales, L. (2013). *Practical Measurement*. Stanford, CA: Carnegie Foundation for the Advancement of Teaching.

Yeager, D. S., Henderson, M., Paunesku, D., Walton, G., Spitzer, B., D’Mello, S., & Duckworth, A. L. (2014). Boring but important: A self-transcendent purpose for learning fosters academic self-regulation. *Journal of Personality and Social Psychology, 107*(4), 559–580. http://dx.doi.org/10.1037/a0037637

Zimmerman, B. J. (2002). Becoming a self-regulated learner: An overview. *Theory into Practice, 41*, 64–70. http://dx.doi.org/10.1207/s15430421tip4102_2

Zimmerman, B. J., Moylan, A., Hudesman, J., White, N., & Flugman, B. (2011). Enhancing self-reflection and mathematics achievement of at-risk urban technical college students. *Psychological Test and Assessment Modeling, 53*, 141–160.