Practical Measurement

David Yeager

University of Texas at Austin

Anthony Bryk

Jane Muhich

Hannah Hausman

Lawrence Morales

Carnegie Foundation for the Advancement of Teaching

Author Note

The authors would like to thank the students, faculty members and colleges who are members of the Carnegie Foundation Networked Improvement Communities and who provided the data used in this report. We would also like to thank Uri Treisman and the University of Texas Dana Center for their input and guidance on the practical theory outlined here, Yphtach Lelkes and Laura Torres for their assistance in creating the practical measures, Peter Jung for his analyses and Angela Duckworth, Christopher Hulleman and Iris Lopez for their comments.

David Yeager is a Fellow of the Carnegie Foundation for the Advancement of Teaching. Address correspondence to David Yeager at 1 University Station #A8000, Austin, Texas 78712 (email: dyeager@utexas.edu).

Abstract

Accelerating the field's capacity to learn *in and through* practice is one key to transforming promising ideas in education into tools, interventions, and professional development initiatives that achieve effectiveness reliably at scale. This paper explains why this type of learning requires a different kind of measurement—measurement that is distinct from the measures commonly used by schools for accountability or by researchers for theory development. The paper presents a theoretical framework for *practical measurement* and illustrates it using a case study of an effort to address the failure rates of community college developmental math students. The paper outlines how a practical theory and set of practical measures were created to assess the causes of "productive persistence"—the set of non-cognitive factors thought to powerfully affect community college developmental math student success. The paper then explains how researchers and practitioners used these measures for practical purposes—specifically, to *assess changes*, *predict* which students were at-risk for course failure, and *set priorities* for improvement work. The paper concludes with a discussion of future directions, including the need for improved behavioral measures and psychometrics tailored for practical measurement.

Practical Measurement

Broadly speaking, the field of educational measurement has evolved to optimize two needs: *accountability* and *theory development*. The former allows us to know with precision how well individual units perform (e.g. districts, schools, classrooms and/or individual students), and the latter allows us to discern in the abstract what might be causing under-performance, and, relatedly, what might alleviate it. Practitioners working with students every day, however, often require an additional kind of information; they want to know how they can reliably improve learning in their classrooms for their particular students, before it is too late. And they need to accomplish this while managing the vast array of demands posed when orchestrating classroom instruction. We argue in this paper that this activity, of learning in and through practice to improve outcomes in the context of everyday practice, often requires a different kind of measurement. We call this *practical measurement*, and in the present article we illustrate why it is necessary, how to create it, and how to use it.

The rest of the article proceeds as follows. First we set the context for practical measurement in recent calls for new ways of thinking about educational change—what has been called *improvement research*. We next explain what improvement research is and why it can assist practitioners to produce changes that are more reliably effective. We then outline why practical measures are essential for conducting improvement research, and discuss why measures created for accountability and theory development, though powerful for different purposes, are often not adequate to inform improvement. Following this general overview, we illustrate in detail the creation and use of practical measures in the context of addressing failure rates in community college developmental math courses. We conclude with a discussion of future directions.

Education's Modus Operandi

Over and over again, change efforts move rapidly across education, with little real knowledge as to how to effect the improvements envisioned by reform advocates (or even whether those improvements are possible). When reformers took aim at the high dropout rates and weak student engagement with high schools, massive effort sprung forth to create new small high schools. Little guidance existed, however, as to exactly how to transform large dysfunctional comprehensive high schools into effective small schools. When reformers focused attention on weaknesses in in-service professional development, a whole new organizational role—the instructional coach—was introduced into schools (Elmore & Burney, 1997; Elmore & Burney, 1998; Fink & Resnick, 2001; Knight, 2007). What coaches actually needed to know and be able to do, and the requisite organizational conditions necessary for them to carry out this work, was left largely unspecified. When reformers recognized the importance of principal leadership, significant investments were directed at intensive principal development programs (Fink & Resnick, 2001). Principals were urged to become instructional leaders even though demands on their time were already excessive and few or no modifications were offered to relieve the latter. The recent introduction of formal teacher evaluation protocols has greatly amplified this stress. When policymakers were unsatisfied with the rate of school improvement, high stakes accountability schemes were introduced. Unintended consequences abounded. The incidence of test-score cheating accelerated and select students were ignored, as accountability schemes directed attention to some students but not others (Jacob & Levitt, 2003; State of Georgia, 2011). The rapid introduction of value-added methods for assessing teachers began well

before the statistical properties and limits of these methods were well understood.[1] Not surprisingly, a host of problems have emerged and political pushback is mounting. Reaching back a bit further, when corporate downsizing was popular, school districts embraced site-based management. The actual domain for such local decision-making however was often left unclear and the necessary resources for carrying out local decisions not provided (Hess, 1995; Bryk, Sebring, Kerbow, Rollow, & Easton, 1998).

In each instance there was a real problem to solve, and in most cases there was at least a nugget of a good reform idea. Educators, however, typically did not know how to execute on these ideas; districts and states lacked the individual expertise and organizational capacity to support these changes at scale; and many policymakers ignored arguably the most important instrument for any of this to work—developing will and agency for engaging these changes among our nation's teachers and principals.

In general, the press to push good ideas into large-scale use rarely delivers on the outcomes promised (Fullan, 2001; Tyack & Cuban, 1995). In some locales a reform might work; in many places, however, it does not. At base is a common story of implementing fast and learning slow. As a field, we *undervalue learning to improve in a way that is systematic and organized*. More specifically, for a change to be successful, educators must learn how to adaptively integrate new materials, processes, and/or roles brought forward by a reform into the organizational dynamics that operate day-to-day in schools (Berwick, 2008; Brown, 1992, Design-Based Research Collaborative, 2003; Bryk 2009; Penuel, et. al. 2011). Assuring efficacy as this adaptive integration occurs, however, is rarely subject to systematic design-development

[1] See reports from the Gates Foundation on the MET study and critical consensus reviews at www.carnegieknowledgenetwork.org

activity. As we will explain, key to achieving the latter are direct measurements of whether the changes being introduced are actually improvements—data that are distinct from the summary evidence routinely used for accountability purposes and also from the measurement protocols used to advance original scientific theories.

Research Focused on Improvement

The central goal of improvement research is for an organization to learn from its own practices to continuously improve.[2] We know from numerous sectors, such as industry and health care, that such inquiries can transform promising change ideas into initiatives that achieve efficacy reliably at scale.

Improvement research taps a natural human bent to learn by doing. This theme about learning in practice has a long tradition reaching back to contributions from both John Dewey (1916) and Kurt Lewin (1935). Informally, learning to improve already occurs in educational organizations. Individual teachers engage in it when they introduce a new practice in their classroom and then examine resulting student work for evidence of positive change. Likewise, school faculty may examine data together on the effectiveness of current practices and share possible improvement ideas. Improvement science seeks to bring analytic discipline to design-development efforts and rigorous protocols for testing improvement ideas. In this way, the "learning by doing" in individual clinical practice can culminate in robust, practical field knowledge (Hiebert, Gallimore, & Stigler, 2002).

[2] The kinds of practical inquiries illustrated are specific examples of "improvement research" i.e. practical disciplined inquiries aimed at educational improvement. The general methodology that guides these individual inquiries is referred to as "improvement science" (Berwick, 2008). For an introduction to this field see Langley et al. (2010).

Several tenets form this activity. The first is that within complex organizations *advancing quality must be integral in day-to-day work* (see, e.g., a discussion of the Toyota Quality Management System in Rother, 2010). While this principle may seem obvious on its face, it actually challenges prevailing educational practice where a select few conduct research, design interventions, and create policies, while vast others do the actual work. Second, improvement research is premised on a realization that education, like many other enterprises, actually has more knowledge, tools, and resources than its institutions routinely use well.[3] The failure of educational systems to *integrate research evidence productively into practice* impedes progress toward making schools and colleges more effective, efficient and personally engaging. Third, improvement science embraces *a design-development ethic*. It places emphasis on learning quickly, at low cost, by systematically using evidence from practice to improve it. A central idea is to make changes rapidly and incrementally, learning from experience while doing so. This is reflected in inquiry protocols such as the plan-do-study-act (PDSA) cycle (Deming, 1986; Imai, 1986; Morris & Hiebert, 2011; Pyzdek & Keller, 2009; Langley et. al., 2009).

Fourth, and anchoring this learning to improve paradigm, is *an explicit systems thinking*—a working theory as to how and why educational systems (and all of their interacting parts) produce the outcomes currently observed. These system understandings generate insights about possible levers for change. This working theory in turn gets tested against evidence from PDSA cycles and consequently is revised over time. The working theory also functions as a scaffold for social knowledge management —it conveys what a profession has learned together about advancing efficacy reliably at scale.

[3] This problem is not peculiar to education and is widespread in different kinds of organizations (see Pfeffer & Sutton, 2000).

Fifth, improvement research is *problem-centered* rather than solution-centered. Inquiries are organized in order to achieve specific measurable targets, not only to spread exciting solutions. Data on progress toward measured targets directs subsequent work. Disciplinary knowledge and methodologies are now used in the service of achieving a practical aim. In the case study we illustrate below, the "core problem" is the extraordinarily high failure rates in developmental mathematics, while the "target" involves tripling student success rates in half the time.

Finally, and arguably most importantly, improvement research maintains a laser-like focus on quality improvement. In this regard, *variability in performance is the core problem to solve.* This means attending to undesirable outcomes, examining the processes generating such outcomes, and targeting change efforts toward greater quality in outcomes for all. This pushes us to look beyond just mean differences among groups, which provides evidence about what *can work.*[4] Instead, the focal concern is whether positive outcomes can be made to occur reliably as new tools, materials, roles and/or routines are taken up by varied professionals seeking to educate diverse sub-groups of students and working under different organizational conditions. *The ability to replicate quality outcomes under diverse conditions is the ultimate goal.*

You Cannot Improve at Scale What You Cannot Measure

[4] To elaborate a bit further, intervention research is typically solution-centered. Such studies seek to demonstrate that some new educational practice or artifact can produce, on average, some desired outcome. The inquiry focus is on acquiring empirical evidence about the practice or artifact. Improvement research draws on such solution-centered inquiries but also reaches beyond this. Its focus is on assembling robust change packages that can reliably produce improvements in targeted problems under diverse organizational conditions, varied sub-groups of students and for different practitioners. While intervention-focused studies seek to make reliable causal inference about what happened in some particular sample of conditions, improvement research aims to assure that measurable improvement in outcomes occur reliably under diverse conditions.

Underlying the tenets of improvement research outlined above is the belief that "you cannot improve at scale what you cannot measure." Hence, conducting improvement research requires thinking about the properties of measures that allow an organization to learn in and through practice. In education, at least three different types of measures are needed, each of which are outlined below. See Table 1.

Measurement for accountability. Global outcome data on problematic concerns—for example, student drop-out rates or pass rates on standardized tests—are needed to understand the scope of the problem and set explicit goals for improvement. These data sources are designed principally to be used as *measures for accountability*. As the name implies, these measures are often used for identifying exemplary or problematic individuals (e.g. districts, schools, teachers) in order to take some specific action, such as extending a reward or imposing some sanction. Because this focus is on measuring individual cases, the psychometrics of accountability data place a high need for reliability at the individual level.

While measures for accountability undoubtedly assess outcomes of interest to policymakers and practitioners, they are limited for making improvements for several reasons. First, the data are typically collected after the end of some cycle (such as the end of the school year), meaning that the people affected by a problematic set of procedures have already been harmed; in a very real sense, the individuals who provide the data (e.g., failed students) will not benefit from the data. Second, because they are global measures of outcomes that are determined by a complex system of forces over a long period of time, the causes that generated these results are often opaque and not tied to specific practices delivered at a specific time. Indeed, a large amount of research on human and animal learning suggests that delayed and causally diffuse feedback is difficult to learn from (see Hattie & Timperley, 2007).

Measurement for theory development. A second and different class of instruments is designed in the course of original academic research. These *measures for theory development* aim to generate data about key theoretical concepts and test hypotheses about the inter-relationship among these concepts. Such measures are also useful in the early stages of designing experimental interventions to demonstrate that, in principle, changing some individual or organizational condition can result in a desired outcome. Such research helps to identify ideas for changes to instruction that might be incorporated into a working theory of practice and its improvement.

In survey research in education, public health, psychology or the social sciences more broadly, measures for theory development often involve administering long, somewhat redundant question batteries assessing multiple small variations on the same concept. For instance, there is a 60-item measure of self-efficacy (Marat, 2005) and a 25-item measure of help-seeking strategies (Karabenick, 2004). By asking a long list of questions, researchers can presumably reduce measurement error due to unreliability and thereby maximize power for testing key relationships of interest among latent variables.

In addition, there is a premium in academic research on novelty, which is often a pre-requisite for publication. Consequently, academic measure development is often concerned about making small distinctions between conceptually overlapping constructs. See for example the six different types of math self-efficacy (Marat, 2005) or seven different types of help-seeking behaviors (Karabenick, 2004). Psychometrically, this leads to a focus on non-shared variance when validating measures through factor analyses and when using predictive models to isolate the relative effects of some variable over and above the effects of other, previously established variables.

All of this is at the heart of good theory development. However, as with measurement for accountability, these types of measures have significant limitations for improvement research. First, long and somewhat redundant measures are simply impractical to administer repeatedly in applied settings such as classrooms. Second, these measures often focus on fine-grained distinctions that do not map easily onto the behaviors or outcomes that practitioners are able to see and act on. Ironically the detail recognized in these academic measures may create a significant cognitive barrier for clinical use. What is the lay practitioner supposed to do, for example, if self-efficacy for *cognitive strategies* is low but self-efficacy for *self-regulated learning* is high, as is possible in some measures of self-efficacy (e.g., Marat, 2005)?

Third, much measurement for theory development in education and the social sciences is not explicitly designed for assessing changes over time or differences between schools—a crucial function of practical measures that guide improvement efforts. One compelling unpublished example comes from research by Angela Duckworth, a leader in the field of measures of non-cognitive factors. She measured levels of self-reported "grit"—or passion and perseverance for long-term goals—among students attending West Point military academy and found that levels of grit actually went *down* significantly over the four years at West Point (Duckworth, personal communication, May 1, 2013), despite the fact that this is highly unlikely to be the case (West Point students undergo tremendous physical and mental challenges as a part of their training). Instead, according to Duckworth, it is likely that they were now comparing themselves to very gritty peers or role models and revising their assessment of themselves accordingly (for empirical, non-anecdotal examples, see Tuttle, Cleason, Knechtel, Nichols-Barrer, & Resch, 2013, or Dobbie & Fryer, 2013). Not that this example does not mean that measures of grit are inadequate for theory development—in fact, individual differences in grit

among students within a school routinely predict important academic outcomes (Duckworth & Carlson, in press; Duckworth, Kirby, Tsukayama, Berstein, & Ericsson, 2010; Duckworth, Peterson, Matthews, & Kelly, 2007). However, such measures may not always be suitable for the purposes of improvement research.

Measurement for Improvement (aka Practical Measurement). Measures for accountability and theory development, although informative for their respective purposes, are insufficient, on their own, for conducting improvement research. The practical work of improvement introduces several new considerations. First, improvement efforts require *direct measurement of intermediary targets* (i.e., "mediators") in order to evaluate ideas for improvement and inform their continued refinement. For example, is a student's motivation and grit actually improving in places where a change has been introduced? And which students benefit most and under what set of circumstances? Second, practical measurement often presses toward *greater specificity* than what occurs with measurement for theory development. Educators need data closely linked to specific work processes and change ideas being introduced in a particular context. Third, increased validity can be gained from measures when *framed in a language targeted* to the *specific units focal for change* (e.g. community college students, many of whom are adults, as compared to elementary school students) and *contextualized around experiences common* to these individuals (e.g. classroom routines used in a course's instruction). Fourth, and most significant from a practical perspective, they need to be *engineered to embed within the constraints of everyday school practice*. For example, a survey given to students routinely in classrooms would need to be brief—in some settings, no more than 3 minutes. These features are described in row 3 of Table 1 and in Table 2.

Uses of Improvement Measures. Practical measures serve several functions. First, they assist educators in *assessing changes*; that is, they can help practitioners learn whether a change that they have introduced is actually an improvement. For this purpose measures need to be sensitive to changes in the short term and quickly accessible to inform subsequent improvement efforts.

A second use for a practical measure is *predictive analytics*. This use answers questions regarding which individuals or groups of individuals are at higher risk for problematic outcomes within a given setting. They can guide educators to target more attention, or supplemental learning supports, in some places rather than others.

A third use for practical measures is *priority setting*. When practitioners are engaged in improvement work, they have to make choices about where best to focus their efforts. Practical measures provide empirical guidance in making these choices. Educators' desire for more equitable student outcomes directs attention toward weakening over time the predictive relationships mentioned above.

We now wish to make these broad themes more concrete by illustrating them in the context of an effort to address a major educational issue. Indeed, a tenet of improvement research is that it is problem-centered, and so when illustrating practical measurement that supports this improvement work we refer to an effort to address a critical educational problem: the outcomes of developmental mathematics students in community college. This effort, carried out in a partnership between the Carnegie Foundation for the Advancement of Teaching and a number of community colleges across the country, embeds improvement research within a network of organizations working more broadly on changes to curriculum and instruction (for initial evidence of efficacy, see Strother, VanCampen, & Grunow, 2013). We believe it is a

helpful case study for imagining how improvement research may be helpful for promoting student success with efficacy and reliability at scale.

A Case Study: Improving Developmental Mathematics Outcomes in Community Colleges

The United States is unique in the world in providing a redemptive path to postsecondary education through community college. Over 14 million students are enrolled in community college, seeking opportunities for a productive career and better life. Community college students are more likely to be low income, the first in their family to attend college, an underrepresented minority and underprepared for college (Bailey, Jenkins & Leinbach, 2005; Rutschow et. al., 2011). Between 60 to 70 percent of incoming community college students typically must take at least one developmental math course before they can enroll in college-credit courses (U.S. Department of Education, 2008; Bailey, Jeong, & Cho, 2010). However, 80 percent of the students who place into developmental mathematics do not complete any college-level course within three years (Bailey, Jeong, & Cho, 2010). Many students spend long periods of time repeating courses and most simply leave college without a credential. As a consequence, millions of people—disproportionally low-income or racial or ethnic minority—each year are not able to progress toward their career and life goals. Equally important, these students lack command of the math that is needed to live in an increasingly quantitative age and to be critically engaged citizens. Developmental math failure rates are a major issue for educational equality and for democracy more generally.

A Pathways Strategy

To address these long-standing challenges, the Carnegie Foundation for the Advancement of Teaching formed a network of community colleges, professional associations, and educational researchers to develop and implement the Community College Pathways program. The program

is organized around two structured pathways, known as Statway® and Quantway®. Rather than a seeming random walk through a maze of possible course options (Zeidenberg & Scott, 2011), students and faculty are now joined in a common, intensive one-semester or year-long experience toward ambitious learning goals, culminating in the awarding of college math credit. Statistics and quantitative reasoning, respectively, are the conceptual organizers for the Pathways. Both Pathways place emphasis on the core mathematics skills needed for work, personal life, and citizenship. The Pathways stress conceptual understanding and the ability to apply it in a variety of contexts and problems. Developmental mathematics objectives are integrated throughout. To date, the Pathways have been implemented in more than 30 colleges in eight states, serving several thousand students (see http://www.carnegiefoundation.org/).

Focusing on "Productive Persistence"

The reasons for the low success rates in developmental math are complex. Developmental math instruction often does not use research-based learning materials or pedagogic practices than can foster deeper learning. Traditional math curricula do relatively little to engage students' interest and demonstrate the relevance of mathematical concepts to everyday life (Carnevale & Desrochers, 2003; also see Hulleman & Harackeiwicz, 2009). Many students have had negative prior math experiences, leading to the belief that "I am not a math person" (Dweck, 2006). These beliefs can trigger anxiety and poor learning strategies when faced with difficult or confusing math problems (Blackwell, Trzesniewski, & Dweck, 2007; Haynes, Perry, Stupinsky & Daniels, 2009; Beilock, Gunderson, Ramirez, & Levine, 2010). This is compounded for some students (e.g., women, African Americans) who are members of groups that have been stereotyped as "not good at math" (Cohen, Garcia, Purdie-Vaughns, Apfel & Brzustoski, 2009; Walton & Spencer, 2009). Research also tells us that students struggle to use the language of

mathematics effectively to understand problem situations, think and reason mathematically, and communicate their learning to others orally and in writing (Gomez, Lozano, Rodela, & Mancervice, 2012; Schoenfeld, 1988).

To respond to these root causes, the Pathways integrate a package of student activities and faculty actions, that aims to increase student motivation, tenacity, and learning skills for success, called *productive persistence*. Productive persistence refers to the behaviors that allow a student to successfully complete their academic training—the *tenacity* to persist, and the *learning strategies* to do so productively.

Prior to our work in this area, the field had not agreed what factors cause productive persistence in community college developmental math or what interventions faculty or course designers might implement to successfully promote it. Instead, ideas come from diverse sources that have not been integrated or made to work successfully across contexts. First, a limited set of precise psychological interventions have had encouraging effects on productive persistence in randomized experiments (for reviews, see Dweck, Walton, & Cohen, 2011; Garcia & Cohen, 2012; Yeager & Walton, 2011), but these have almost never been tested with community college students, and, at least in the published literature, have only been tested at a small scale (Yeager, Paunesku, Walton, & Dweck, 2013). Furthermore, it is not known how to integrate the overall findings from these experiments into the daily practices engaged in by practitioners. Second, there are many comprehensive interventions that have been the subject of high quality evaluations—such as "learning communities" (Weiss, Visher, Wathington, Teres, & Schneider, 2010), intensive mentoring, (Visher, Butcher, & Cerna, 2010) or comprehensive student success courses (Rutschow, Cullinan, & Welbeck, 2012). Discouragingly, each of these evaluations has shown small or no effects on student performance or credit attainment beyond the treatment

period. Finally, there are also many clinical rules of thumb about effective student engagement tactics —instincts formed through experience and common among wise practitioners—but these lack an evidentiary basis for effectiveness. How can practitioners distill the ideas from these diverse sources into a manageable set of practices that will promote student success in community college classrooms? And how can these practices be enacted *reliably, at scale, by diverse practitioners working in diverse settings?* To learn how to do this it is important to first agree on a common framework and a common set of measures to inform improvement efforts.

A Practical Theory to Guide Improvement Work

A tenet of improvement research is an explicit systems thinking. Thus, beginning improvement research requires the development of a "practical theory"--an easily-interpretable conceptual framework--that practitioners can see as useful in guiding their work, while remaining anchored in the best available empirical research. We explain and illustrate this next.

What is a Practical Theory for Improvement?

We begin by noting that a practical theory is not a *disciplinary* theory, in that it does not seek to document novel features of human psychology or social or economic processes that shape the ways humans in general think or behave. Instead, a practical theory draws on both the wisdom of practice as well as insights from academic theories to guide practice improvement. While disciplinary theories emphasize novelty, counter-intuitiveness or fine distinctions—and as a result have a highly important role in science—a practical theory uses only those distinctions or novel ideas that can reliably motivate practitioner action in diverse contexts. A practical theory is also not a *general educational* theory. It is not designed to be an account of all relevant problems (for instance, motivation among students of all levels of ability or of all ages). Rather, in the present case, it was co-created with researchers at the Carnegie Foundation and practitioners in

community colleges and was tailored for the challenges faced specifically by developmental math students. Nevertheless, because creating a practical theory is an activity activity anchored in disciplinary research, doing this might insights both for subsequent disciplinary inquiries and for practical theories of other educational problems.

To re-iterate, the virtue of a practical theory is not that it is new or non-obvious or exhaustive. To the contrary, the virtue of a practical theory is that each element is immediately recognizable to both practical experts and theoretical experts, each of whom deeply understands the problem of practice, through their own lenses. Such theories function as a useful guide for practice improvement while remaining grounded in current scientific knowledge.

Steps for Creating a Practical Theory

How can one create practical theory? In the case of productive persistence, we began with the assumption that much good work had already been done in both research and practice. We therefore started by seeing whether a framework could be created rapidly, in just 90 days, by drawing on the expertise already present in the field. To do so we conducted a "90-day inquiry cycle" (Institute for Healthcare Improvement, 2010; Huston, & Sakkab, 2006) that scanned what was known in the field about the conceptual area. We cast a wide net to generate an initial list of concepts that might be related to productive persistence. Constructs, measures, theories, and interventions were found through conversations with academic experts, surveys of and interviews with practitioners, as well as keyword searches in the leading databases (i.e., Google Scholar, PsychInfo, etc.). We identified over 1000 possible constructs. Such a list, however, is impractical. Therefore, using the process outlined below, we reduced the list to five broad areas with a handful of specific elements within each. These were:

- **Skills, habits and know-how to succeed in a college setting**, such as coping with math anxiety, using effective study strategies, or having self-discipline when managing competing goals, desires, and commitments.
- **Students believe they are capable of learning math,** including fixed vs. growth mindsets about potential to learn and improve in math.
- **Students believe the course has value**, including judgments that the coursework has relevance for degree completion or for important personal and social goals.
- **Students feel socially tied to peers, faculty and the course**, including a feeling of connection to the course, experiences of stereotype threat or uncertainty about one's social belonging in the setting.
- **Faculty support students' mindsets and skills**, including instructors' beliefs in students' potential to improve at math or instructors' skills at promoting engagement.

We accomplished this reduced list by applying two broad filters:

Filter 1: Does the state of the science support the general importance of the concept? As a first pass to reduce this initial list to a manageable size, we relied heavily on the published scientific record. We asked: (a) Are there data (ideally from experiments) supporting a *causal* interpretation for the concept? (b) Is the concept distinct enough to yield practically distinctive implications (not self-efficacy for cognitive strategies vs. self-efficacy for self-regulated learning) and was it theoretically precise enough to be useful (i.e., not just "feeling connected" to the classroom)?; (c) Is the concept an underlying cause or better viewed as a mediator of some concept that causally precedes it (i.e., although self-efficacy is highly predictive and important, the practical theory focused on the causal antecedents of self-efficacy, such as a fixed vs. growth

mindset, Dweck, 1999)? This eliminated a great many of the possible constructs and led to the re-framing of many of those that remained.

Filter 2: Does the science suggest that this concept is likely to be relevant *for improvement in this specific context*? The second filter was more extensive and fine-grained than the first and refined the framework to be more specifically useful for the given improvement context. It involved answering the following questions: (a) Is the concept likely to be amenable to change via the *systems of influence in place in the improvement setting* (i.e., either by a faculty member's behaviors or by the structure of the course)? For instance, being responsible for dependents is likely a cause of low performance for some college students, but this factor is not amenable to change through efforts of a faculty member or college; (b) Is the concept likely to be amenable to change *within the time duration of instructional setting*? For instance, the personality trait of conscientiousness (Duckworth, Weir, Tsukayama, & Kwok, 2012; Eisenberg, Duckworth, Spinrad, & Valiente, in press) might be highly predictive of achievement, but, at least so far, there is little or no evidence that this trait is malleable in the short term or that existing measures of the trait are sensitive to short-term changes; (c) Is the concept likely to be *measured efficiently in practical settings*? For instance, executive function and IQ are strong predictors of math performance (e.g., Clark, Pritchard, & Woodward, 2010; Mazzocco & Kover, 2007) but valid assessments are, at least currently, time- and resource-intensive and impractical for repeated measurement by practitioners; and (d) Are there known or suspected moderators that suggest the factor may matter less for the population of interest, and hence may provide less leverage as a focus for improvement?

Finalizing the practical theory. After applying these two filters, an initial framework for productive persistence was created. The model was then "tested" and refined by using focus

groups and conversations with faculty, researchers, college counselors and students. In these "testing" conversations, practitioners opined (a) whether or not they felt that the framework captured important influences on developmental math achievement (i.e. face validity); and (b) whether the concepts composing the framework were described in a way that made them understandable and conceptually distinct. This led to a number of cycles of revision and improvement of the framework.

After some initial use in work with community college faculty, the framework was "tested" again in January 2012 via discussions at a convening of expert practitioners and psychologists hosted at the Carnegie Foundation.[5] The product of this effort, still a work in progress, is depicted in Figure 2.

Formulating a Practical Measure

A practical theory allows researchers to work with practitioners on an agreed-upon set of high-leverage factors thought to influence an outcome of interest. But, as we have been suggesting, using the practical framework requires implementing practical measures of the factors described in it. In the present case, after identifying and refining the five conceptual areas relevant to productive persistence (Figure 1), a next step was to create a set of practical measures to assess each. Because many of the ideas in the concept map had come from the academic literature, there were measures available for each. A comprehensive scan of the field located roughly 900 different potential survey measures.

[5] The meeting in which the practical theory was vetted involved a number of the disciplinary experts whose work directly informed the construct in the framework; these were Drs. Carol Dweck, Sian Beilock, Geoffrey Cohen, Deborah Stipek, Gregory Walton, Christopher Hulleman, and Jeremy Jamieson, in addition to the authors.

By and large, however, available measures failed the test of practicality. Many items were redundant, theoretically-diffuse, double-barreled questions using vocabulary that would be confusing for respondents learning English or with low cognitive ability or levels of education. In addition, evidence of predictive validity, a primary criterion for a practical measure, was rare. For instance, an excellent review of existing non-cognitive measures (Atkins-Burnett, S, Fernandez, C, Jacobson, J, & Smither-Wulsin, C., 2012; for a similar review see U.S. Department of Education, 2011), located 196 survey instruments coming from 48 independent empirical articles. Our team of coders reviewed each of these and could not locate any objective validity evidence (i.e., correlations with test scores or official grades) for 94% of measures. Administration in community college populations was even more rare; our team could find only one paper that measured the concepts identified in our practical theory and showed relations to objective course performance metrics among developmental mathematics students. Of course, many of these measures were not designed for improvement research; they were designed to test theory and as such were often validated by administering them to large samples of captive undergraduates at selective universities. Practical measurement, by contrast, has different purposes and therefore requires new measures and different methods for validating them.

Another key dimension of practicality is brevity. In the case of the Community College Pathways project, faculty agreed to give up no more than 3 minutes for survey questions. This created a target of approximately 25 survey items that could be used to assess the major constructs in Figure 1 and also serve each of the purposes of practical measurement (assessing changes, predictive analytics, and setting priorities). Therefore, our team took the list of 900 individual survey items and reduced them to roughly 26 items that, in field tests with community college students, took an average of 3 minutes to answer.

How was this done? At a high level, we began by organizing items into clusters that matched the broad conceptual areas shown in Figure 1. Many items were overlapping or nearly identical. This step reduced large numbers of items. Next, we were guided by theory in selecting sub-sets of items that matched experimental operationalizations. This too eliminated large numbers of items. Next we selected items that followed principles of optimal item design. When such items were not found, then items were re-written. In addition, redundant sets of items were reduced to one item or to small clusters of 3-4 items assessing distinct components of a broader concept in the diagram. We explain these steps in greater detail below.

Step 1: Guided by theory. The process of creating the practical measures began by looking to the experimental literature to learn what effectively promotes tenacity and the use of effective learning strategies, the hallmarks of productive persistence. We then selected or re-wrote items so that they tapped more precisely into the causal theory. For instance, while an enormous amount of important correlational research has focused on the impact of social connections for motivation, (e.g., Wentzel & Wigfield, 1998) some experimental literature focuses more precisely on a concept called "belonging uncertainty" as a cause of academic outcomes in college (Walton & Cohen, 2007; 2011). Walton and Cohen's (2011) theory is that if a person questions whether they belong in a class or a college, it can be difficult to fully commit to the behaviors that may be necessary to succeed, such as joining study groups or asking professors for help. Of significance to practical measurement, it has been demonstrated that an experimental intervention alleviating belonging uncertainty can mitigate the negative effects associated with this mindset (Walton & Cohen, 2011). Such experimental findings provide a basis for item reduction. Instead of asking students a large number of overlapping items about liking the school, enjoying the school, or fitting in at the school, our practical measure asked a

single question: “When thinking of your math class, how often, if ever, do you wonder: Maybe I don’t belong here?” As will be shown below, this single item is an excellent predictor of course completion and course passing (among those who completed), and this replicates in large samples, across colleges and Pathways (Statway or Quantway).

A similar process was repeated for each of the concepts in the practical theory. That is, we looked to the experimental literature for methods to promote relevance (Hulleman & Harackeiwicz, 2009), supporting autonomy (Vansteenkiste et al., 2006), a “growth mindset” about academic ability (Blackwell, Trzesniewski, & Dweck, 2007), goal-setting and self-discipline (Duckworth & Carlson, in press; Duckworth, Kirby, Gollwitzer, & Oettingen, in press), skills for regulating anxiety and emotional arousal (Ramirez & Beilock, 2011; Jamieson et al., 2010), and others. We then found and re-wrote items that were face-valid and precisely related to factors that were malleable and high-leverage, allowing for fewer but more precise measures.

Step 2: Optimal item design. In addition to selecting theoretically-precise items, we revised the wording of the items according to optimal survey design principles so as to maximize information from very few questions (see Krosnick, 1999; Schumann & Presser, 1981). In fact, there is a large experimental literature in cognitive and social psychology that has created practical measures in a different setting: measuring political attitudes over the phone in national surveys (Krosnick, 1999; Krosnick & Fabrigar, in press; Schumann & Presser, 1981; Tourangeau, Rips, & Rasinski, 2000). Unlike much measurement for theory development in psychology and education, public opinion surveys must be face valid enough to withstand accusations of bias from the media and the lay public. But they must also be brief and clear. And verbal administration can exaggerate the differences in measurement accuracy among low-

education respondents (Holbrook, Krosnick, Moore, & Tourangeau, 2007; Krosnick & Alwin, 1987). Therefore a large number of national experiments have discovered how to maximize accuracy for low-education sub-groups in particular (Narayan & Krosnick, 1996; see Krosnick, 1999). Such findings are relevant for administration to students taking developmental math in community college because they are, by definition, low-education respondents.

Which lessons from the public opinion questionnaire design literature were relevant? One strong recommendation is to, whenever possible, avoid items that could produce acquiescence response bias (Krosnick & Fabrigar, in press). Acquiescence response bias is the tendency for respondents to "agree", say "yes" or say "true" for any statement, regardless of its content (Saris, Revilla, Krosnick, & Shaeffer, 2010; Schumann & Presser, 1981). For example, past experiments have found that over 60% of respondents would agree with both a statement and its logical opposite (Schumann & Presser, 1981). Such a tendency can be especially great among low-education respondents (see Krosnick, 1991), which, again, were the targets of our measures. Therefore, unless we otherwise had evidence that a given construct was best measured using an agree / disagree rating scale (as happened to be the case for the "growth mindset" items, Dweck, 1999),[6] we wrote what are called "construct specific" items.

What is a "construct specific" question? An item asking about math and statistics anxiety, for example, could be written in agree / disagree format as "I would feel anxious taking a math or statistics test" (Response options: 1 = *Strongly disagree;* 5 = *Strongly agree*) or it could be written in construct specific format, as in "How anxious would you feel taking a math

[6] Surprisingly, in pilot experiments, the traditional agree / disagree fixed mindset questionnaire items (Dweck, 1999) showed improved or identical predictive validity compared to construct-specific questions, the only such case we know of showing this trend (cf. Saris, Revilla, Krosnick, & Shaeffer, 2010; Schumann & Presser, 1981).

or statistics test?" (Response options: 1 = *Not at all anxious;* 5 = *Extremely anxious*). In fact, we tested these two response formats. We conducted an large-sample ($N > 1{,}000$) experiment that randomly assigned developmental math students to answer a series of items that assessed anxiety by using either agree / disagree or construct-specific formats, similar to those noted above. This was done during the first few weeks of a course. We then assessed which version of these items was more valid by examining the correlations of each with objective behavioral outcomes: performance on an assessment of background math knowledge at the beginning of the course and performance on the end of term comprehensive exam, roughly three months later. We found that the construct-specific items significantly correlated with the background exam, $r = .21$, $p < .05$, and with the end-of-term exam, $r = .25$, $p = < .01$, while the agree / disagree items did not, rs = .06 and .09, *n.s.*, respectively (and these correlations differed from one another, ps < .05), demonstrating significantly lower validity for agree / disagree items compared to construct-specific items.

We employed a number of additional "best practices" for reducing response errors among low-education respondents. These included: fully stating one viewpoint and then briefly acknowledging the second viewpoint when presenting mutually exclusive response options (a technique known as "minimal balancing;" Schaeffer, Krosnick, Langer, & Merkle, 2005); using web administration, because laboratory experiments show that response quality is greater over the web (Chang & Krosnick, 2010); displaying response options vertically rather than horizontally to avoid handedness bias in primacy effects (Kim, Krosnick, & Cassanto, 2012); ordering response options in conversationally natural orders (Holbrook, Krosnick, Carson, and Mitchell, 2000; Tourangeau, Couper, & Conrad, 2004); and asking about potentially sensitive

topics using "direct" questions rather than prefacing them with "some people think… but other people think…" (Yeager & Krosnick, 2011; 2012) in addition to others.

Step 3: Contextualizing and pre-testing. After an initial period of item writing, the survey items next went through a process of customization to the perspectives of community college practitioners and students. Following best practices, we also conducted cognitive pre-tests (Presser, Couper, Lessler, Martin, Martin, Rothgeb, & Singer. 2004) with current developmental math students to surface ambiguities or equivocations in the language. We paid special attention to how the items may have confused the lowest performing students or students with poor English skills—both groups that would be especially likely to under-perform in developmental math, and therefore groups that ideally the practical measures would help us learn the most about how to help. This led to re-writing of a number of items, and also confirmation that many survey items were successfully eliciting the type of thinking they were designed to elicit.

Step 4: Finalizing the resulting practical measure. These efforts to produce a "practical" self-report measure of productive persistence resulted in 26 items. In their subsequent use in the Pathways, however, not all of these items proved to be predictive of student outcomes, either on an individual level or on a classroom level. When the underlying construct involved several distinct but correlated thoughts or experiences, items were designed to be combined into small clusters (no more than 4 items; and in such cases one item was written for each distinct thought or experience and the combined into the higher-level construct). Altogether, 15 survey items were used to measure the following 5 constructs (see the online supplement for exact wording and response options):

- **Math anxiety,** 4 items (e.g., "How anxious would you feel the moment before you got a math or statistics test back?").
- **Mindsets about academic potential,** 4 items (e.g., "Being a 'math person' or not is something about you that you really can't change. Some people are good at math and other people aren't).
- **Mindsets about the value of the coursework**, 3 items (e.g., "In general, how relevant to you are the things that are taught in math or statistics class?).
- **Mindsets about social belonging**, 3 items to assess social ties (e.g., in addition to the belonging uncertainty measure noted above, "How much do you think your professor would care whether you succeed or failed in your math or statistics class?"), and 1 item to assess stereotype threat**,** ("Do you think other people at your school would be surprised or not surprised if you or people like you succeeded in school?").
- **"Grit"**: As a behavioral indicator of "grit" (Duckworth et al., 2007), we used whether a student answered every question on a background math test.

It is important to note that while these items provide a promising example of the potential for practical measures, in every case both the construction and use of the measures could be further improved. For instance, while these each measure aspects of the practical theory in Figure 1, some measures that we created did not show meaningful validity correlations. And so further development is needed to more fully measure all of the concepts in the practical theory. Nevertheless, the resulting practical measure is useful for illustrating the uses of practical measures, as we demonstrate below.

Step 5: Use in an instructional system. After this process and some initial piloting, the brief set of measures was embedded in the Pathways online instructional system—a website

hosting students' textbooks and homework. After logging in, students were automatically directed to complete the items before completing their homework online, both on the first day of class and again four weeks into the course. In this way, causes of students' productive persistence could be assessed efficiently and practically, without effort from faculty, and with response rates comparable to government surveys in many cases (for exact response rates, see the online supplement).

Illustrative Examples Using Practical Measurement to Improve

As noted earlier, practical measurement is helpful for (1) assessing changes, (2) predictive analytics and (3) priority setting. We illustrate each of these below in the context of our case study and summarize key differences in Table 3.

1: Assessing Change

One use for practical measures is to assess whether changes implemented were, in fact, improvements—at least in terms of the concepts outlined in the practical theory. An assumption in improvement research is that variability in local practice a will be linked to variability in student outcomes. The challenge for improvement researchers is to measure both of these so as to learn how to change practice in ways that reduce variability in performance and create quality outcomes for all.

Evaluating a "Starting Strong" package. As noted, both practitioner accounts and empirical studies find that the first few weeks of the term are a critical period for student engagement. When students draw early conclusions that they cannot do the work or that they do not belong then they may withhold the effort that is required to have success in the long term, starting a negative recursive cycle that ends in either course withdrawal or failure (Cook, Purdie-Vaughns, Garcia, & Cohen, 2012). Similarly, in the first few class periods students join or do not

join study groups that will ultimately be informal networks for sharing tips for course success. After a brief period of malleability, informal student networks can be remarkably stable and exclusive over the course of the term and also strikingly predictive of student learning over time (Vaquero & Cebrian, 2013). The productive persistence conceptual framework posits that if faculty successfully created a classroom climate that helped students see their academic futures as more hopeful and that facilitated developing strong social ties to peers and to the course, students may gradually put forth more effort and, seeing themselves do better, might show an upward trajectory of learning and engagement.[7]

In light of these possibilities, the productive persistence activities consisted of classroom routines in the form of a "Starting Strong" package. This consisted of a set of classroom routines timed for the first few weeks of the term and targeted toward the major concepts in the conceptual framework (Figure 1): reducing anxiety, increasing interest in the course, forming supportive students' social networks, etc.. For example, the "Starting Strong" package included a brief, one-time "growth mindset" reading and writing activity that had been shown in some past experimental research to increase overall math grades among community college students (see Yeager et al., 2013; cf. Blackwell et al., 2007). There were also classroom activities for forming small groups, getting to other students in the class, etc.

Were the practical measures effective at assessing changes? As a first look, we examined the productive persistence survey on the first day of class and after three weeks of instruction. Evidence on the efficacy of the Productive Persistence "Starting Strong" package, presented in Figure 3, was encouraging. The results, presented in standardized effect sizes, show moderate to large changes in four measured student mindsets after the first three weeks of exposure to

[7] For a psychological analysis, see Garcia and Cohen (2012)

Statway. As instruction began, students' interest in math increased, their beliefs about whether math ability is a fixed quantity decreased, math anxiety decreased, as did their uncertainty about belonging. However, these effects did not occur in every college and for every sub-group of students; the latter results, in conjunction with predictive validity findings (see below), informed subsequent improvement priority setting (below).

2: Predictive Analytics

At-riskness index. Another use for practical measures is to assess whether data collected on the first day of class might be predictive of a student's probability of successfully completing the course. For this purpose, we developed an "at-riskness" indicator based on student responses to the productive persistence questions asked on the first day of the course. This type of measure might support quality improvement because early interventions, tailored to student needs and delivered by faculty, might increase the likelihood of success for students at risk for failure.

Data from three of the main concepts shown in Figure 1 were used to form the at-riskness indicator: (1) Skills and habits for succeeding in college; (2) Students believe they are capable of learning math; and (3) Students feel socially tied to peers, faculty and course of study. Data on the perceived value of the course were not included in at-riskness indicator because on the first day of the course students would not be expected to provide meaningful information about how interesting or relevant they found it. The measures about faculty's mindsets and skills were also not the focus of the at-riskness index because, in the current analysis, our objective was to understand variance in *student* risk factors *within* classrooms, not risk factors at the teacher level (the latter is presented next).

Analyses empirically derived cut points that signaled problematic versus non-problematic responses on five different risk factors for the three concepts listed above (anxiety, mindsets

about academic ability, social ties, stereotype threat, and "grit"). The systematic procedure for doing this is presented in the online supplemental material. Analyses then summed the number of at-risk factors to form an overall at-riskness score ranging from 0 to 5.

As illustrated in Figure 4, productive persistence risk level showed a striking relation to course outcomes (also see the online appendix). Students with high risk on day 1 were roughly twice as likely to fail an end-of-term exam several months later as compared to low-risk classmates. Testifying to the robustness of these findings, these findings replicated in both the Statway colleges and the Quantway colleges, totaling over 30 institutions. Furthermore, the productive persistence at-riskness index from the first day of the course predicted end-of-term exam performance even when controlling for mathematical background knowledge and student demographic characteristics such as race or number of dependents at home (see the online appendix for hierarchical linear models). Thus, by following the procedure noted above for creating a practical theory and practical measures, a set of questions that takes less than 3 minutes to administer can identify, on day 1, students with a very low chance of successfully completing the course.

Real-time student engagement data. The analyses above show that it is possible to identify *students* with higher levels of risk for not productively persisting. But is it possible to identify *classes* that either are or are not on the path to having high rates of success? If it were possible, for example, to capture declines in feelings of engagement before they turned into course failures, interventions might be developed to help instructors keep students engaged.

As a first step toward doing this, the Carnegie Foundation instituted very brief (3-5 question) "pulse check" surveys in the online instructional system—the website Statway students use to access their textbook and do their homework. Every few days, after students logged in, but

before they could visit the course content, students were redirected to a single-page, optional survey consisting of three to five items. Students were asked their views about the course content (e.g. whether there were language issues, whether it was interesting and relevant), but, most crucially for the present purposes, they were asked "Overall, how do you feel about the Statway course right now?" (*Extremely negative* = 1, *Mostly negative* = 2, *Mostly positive* = 3, *Extremely positive* = 4). As shown in Figure 4, nearly all classrooms' began with high levels of enthusiasm. But this cooled over time toward more realistic level of being "mostly positive" on average. What differentiated classes with high pass rates (80% or more) from those with low pass rates (less than 80%), however, was what happened after that initial decline in enthusiasm. Successful classrooms slowed and even reversed the negative trend in student reports. By contrast, less successful classes showed a continued downward decline, with students becoming more negative as the term continued to progress.

Thus, with only a single item, asked routinely via a homework platform, we could obtain real-time data that differentiated among classes in their ultimate success rates several months later. If future analyses replicated these trends across contexts, it is easy to see how this practical measure could constitute an effective early warning system for targeting classroom-level improvement efforts, such as targeted professional development to teachers.

3: Priority Setting

As noted, a third important use of data when conducting improvement research is to assess which aspects of a practical theory, to date, have not yet been successfully addressed. For instance, we found that one survey item administered in the fourth week of the course, assessing belonging uncertainty (Walton & Cohen, 2007; 2011), was the single best predictor of whether students dropped the course before the end of the semester, even after controlling for background

math knowledge and demographic-personal characteristics such as race/ethnicity, income, number of dependents in the home, and number of hours worked (Figure 5). Furthermore, among students who did not withdraw from the course, this item was an excellent predictor of whether students met the minimum threshold for being prepared for subsequent math coursework (a B- or better; Figure 5).

These findings have led directly to network priorities for improvement efforts to address belonging uncertainty. These results were a signal to faculty that in their classes belonging uncertainty was not being sufficiently addressed, and mattered a great deal for their students. We have learned that this kind of "local empiricism" can powerfully motivate faculty improvement efforts. Indeed, several efforts have emerged in the network to address this priority. Faculty are now collaborating with academic researchers in an effort to adapt to the community college the context an experimental social-psychological intervention that has demonstrated effect in this area (Walton & Cohen, 2011).

In addition, faculty are testing a set of new classroom routines developed specifically to enhance students' social connections in class. Faculty have begun to conduct plan-do-study-act cycles (Deming, 1986; Imai, 1986; Morris & Hiebert, 2011; Pyzdek & Keller, 2009) on new routines for creating a sense of social belonging on a daily basis in their classrooms. These routines focus on seemingly-mundane changes to procedures that nevertheless might affect students' feelings of connection to the course—for instance, routines for emailing absent students or improved routines for creating and maintaining collaborative small-groups. Faculty track practical measures of behaviors—like attendance—and also periodically administer the survey items assessing mindsets about social belonging. The goal of this improvement activity is to implement a change, measure its intended consequences, look at one's data, and then adjust—

all while students are still in a course, before they have withdrawn or failed. Ultimately, each term faculty will be able to conduct many such cycles of improvement across the network for other concepts in the practical theory outlined in Figure 1, ideally leading to accelerated and reliable improvements in student outcomes at scale.

Implications for Efforts to Scale Productive Persistence

Education reformers are rightfully enthusiastic about the potential for research on productive persistence to contribute to the improvement of student outcomes (see Dweck et al., 2011; Farrington et al., 2012). Our emphasis on creating networks engaged in improvement research related to productive persistence is not based on shortcomings in the evidence in support of the mindsets but rather on the observation that there have been *many* promising ideas in the history of education, and, shockingly, very *few* examples of one that has been successfully been implemented with reliability at scale.

We have proposed that improvement research can be a helpful way forward. We have shown that it is possible to develop an understandable "practical" framework for the broad and seemingly incoherent field of student success / non-cognitive factors–what we have titled productive persistence. Next, we have shown that it is possible to adapt measures that were originally developed for theory development and use them in the context of improvement work on a new developmental math course. Importantly, even brief but fine-tuned measures could be highly predictive of course outcomes and help a network engaged in improvement research assess changes, identify students at risk, and set improvement priorities.

Finally, we note that this work is only beginning. In the future, we expect that additional disciplinary research will further refine theories and contribute additional interventions to more fully address the barriers contributing to underperformance. And through the work of educators

using practical measures to conduct improvement research, it will be possible to implement the strong ideas coming out of laboratory research, and also generate new, related practices stemming from faculty wisdom so as to promote productive persistence. While this work is nascent, we hope it provides some guidance for how the field might learn more quickly in and through practice.

Future Directions for Research on Practical Measurement

Behavioral Assessments

In the present article we have illustrated the development and use of predominatly self-report practical measures of productive persistence. We have done this because a great deal of past research (Krosnick, 1999) has supported the overall assumption that if you ask people sensible questions that they know the answer to, under circumstances where they feel able to report their true opinions, then you can gather highly predictive data from even brief sets of questions.

However in many cases it would also be desirable to develop behavioral indicators of the concepts outlined in a practical theory to supplement these self-reports. This is true in part because of reference bias (Biernat, 2003), which is the tendency for a self-report measure to rely on the subjective frame of reference of the respondent (for instances where reference bias may have occurred in the assessment of non-cognitive factors, see Tuttle et al., 2013, or Dobbie & Fryer, 2013).

To bypass some of the potential limitations of self-reports, one could design novel behavioral measures. For instance, one could analyze whether students review their work or not when doing homework in an online course management platform. To determine whether a classroom has successfully created a challenge-seeking culture in the first few weeks of the

course, a person conducting improvement research could embed opportunities for student choice in the level of difficulty of tasks and then assess the percent of students who chose hard tasks (where they could learn a lot) as opposed to easy tasks (where they could get a high score; for examples of such measures, see Mueller & Dweck, 1998). And to determine whether students have developed productive study habits, a practitioner could track the percent of students who reviewed past problems or online textbook content before attempting new hard problems. Indeed, recent research has pointed to the surprising power of the behavioral residue of completing assignments—for instance, whether students complete all of the assigned problems--to indicate non-cognitive factors such as grit or self-control (see the behavioral measure of "grit" above, in the at-riskness index; also see Hedengren & Stratmann, 2012). More generally, behavioral indicators might be unobtrusively added to online learning environments and collected and reported on automatically, making them highly practical. Even simple behaviors such as even logging into an online platform, clicking through problems versus honestly attempting them, etc., might be a rich source of "non-cognitive" data that can inform improvement efforts.

Improved Psychometrics

A second future direction involves psychometrics. Much psychometric theory has been developed to optimize measures for accountability or theory development. As outlined in Table 1, one of the primary psychometric criteria for accountability measures is high reliability at the level at which rewards or punishments are being delivered, to avoid both false negatives and false positives that unfairly affect a teacher, a school, or a district. Some of the primary psychometric criteria for effective measures for theory development are internal consistency, reliability, and construct validity. Each of these types of measures come with relevant summary

statistics that researchers can readily interpret to ascertain the likely suitability of a measure either for accountability or theory (e.g., Cronbach's alpha, model fit in a confirmatory factor analysis, etc.).

Because the purposes for practical measures are distinct, it is also worth considering whether psychometricians might develop or adapt new summary statistics that are more helpful for indicating the suitability of items or clusters of items in a practical measure. For instance, ideally a practical measure for purposes of predictive analytics could not be evaluated using internal consistency reliability (because it would involve one item or one behavior, or because it would involve small clusters of items that were designed to be non-overlapping but modestly correlated). Items measuring different constructs would also ideally have no loading on a common factor (because there would be no redundant measures or clusters of items in the battery). By contrast, predictive validity is at a premium—does this measure predict long-term outcomes of interest?—as is the potential for the measure to be sensitive to even small changes in instruction or classroom culture—does this measure tell me whether, this week, I have successfully addressed the leading indicators of course performance? Thus, in the absence of effective intervention, a practical measure would be strongly predictive of outcomes, but in the presence of the intervention, the practical measure's validity would be driven to zero (because the risk factor had been successfully addressed). In some cases, practical measures could be evaluated in terms of how well change scores operate as predictors of long-term course outcomes. Thus, a different or revised set of rules would be needed for reviewers to evaluate manuscripts or grant proposals that include practical measures, and a different set of guidelines is necessary for improvement researchers or practitioners to select or create appropriate practical

measures. Clearly, much more theoretical work, simulation work, and field research are required to do this sufficiently.

Conclusion

We have argued that educators need to be able to assess whether the instruction they deliver in a classroom is, in fact, leading to the changes they hope for, in real time, well before students become academic casualties. Although measurement for accountability is important for signaling a problem, relying on such measures for improvement is analogous to standing at the end of the production process and counting the number of broken widgets. The quality of the end product is an aggregate consequence of many discrete processes that operate within a complex production system. Quality improvement entails deeper information about system processes, where undesirable outcomes stem from, and targeting subsequent improvement based on this knowledge. Seeking to remediate the problem at the end of the line is neither an effective nor efficient solution (Rother, 2010).

Educators need both more frequent data and also different kinds of information than they normally get—measures that can help them improve their actual practices. We look forward to future research on methods to create and embed practical measures in networks of research and practitioners engaged in improvement research. We believe this can play a substantial role in the quality improvement of educational processes at scale.

References

Atkins-Burnett, S, Fernandez, C, Jacobson, J, & Smither-Wulsin, C. (2012). *Landscape analysis of non-cognitive measures*. Princeton, NJ: Mathematica Policy Research.

Bailey, T., Jenkins, D., & Leinbach, T. (2005). *What we know about community college low-income and minority student outcomes: Descriptive statistics from national surveys.* New York, NY: Teachers College Community College Research Center, Columbia University. Retrieved May 13, 2013 from http://www.eric.ed.gov/PDFS/ED484354.pdf

Bailey, T., Jeong, D. W., & Cho, S. W. (2010). Referral, enrollment, and completion in developmental education sequences in community colleges. *Economics of Education Review*, 29, 255-270.

Beilock, S. L., Gunderson, E. A., Ramirez, G., & Levine, S. C. (2010). Female teachers' math anxiety affects girls' math achievement. *Proceedings of the National Academy of Sciences, USA, 107*, 1060-1063.

Berwick, D. M. (2008). The science of improvement. *The Journal of the American Medical Association, 299*, 1182-1184.

Biernat, M. (2003). Toward a broader view of social stereotyping. *American Psychologist, 58*, 1019.

Blackwell, L. S., Trzesniewski, K. H. & 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.

Brown, A. L. (1992). Design Experiments: Theoretical and Methodological Challenges in Creating Complex Interventions in Classroom Settings. *Journal of the Learning Sciences, 2*, 141–178. doi:10.1207/s15327809jls0202_2

Bryk, A. S. (2009). Support a science of performance improvement. *Phi Delta Kappan, 90*, 597-600.

Bryk, A., Sebring, P. B., Kerbow, D., Rollow, S., & Easton, J. (1998). *Charting Chicago school reform: Democratic localism as a lever for change*. Boulder, CO: Westview Books.

Carnavale, A. and Desrochers, D. (2003). *Standards for what?: The economic roots of K-16 reform*. Princeton, NJ: Educational Testing Service.

Chang, L. & Krosnick, J. A. (2010). Comparing oral interviewing with self-administered computerized questionnaires: An experiment. *Public Opinion Quarterly, 74*, 154-167.

Clark, C. A., Pritchard, V. E., & Woodward, L. J. (2010). Preschool executive functioning abilities predict early mathematics achievement. *Developmental psychology, 46*, 1176.

Cohen, G. L., Garcia, J., Purdie-Vaughns, V., Apfel, N., & Brzustoski, P. (2009). Recursive processes in self-affirmation: Intervening to close the minority achievement gap. *Science, 324*, 400-403.

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.

Deming, W. E. (1986). *Out of the Crisis, 1986*. Cambridge, Mass.: Massachusetts Institute of Technology Center for Advanced Engineering Study. xiii, 507.

Dewey, J. (1916). *Democracy and education: An introduction to the philosophy of education.* New York, NY: Macmillan.

Dobbie, W., & Fryer, R. (2013). The medium-term impacts of high-achieving charter schools on non-test score outcomes. Retrieved August 8, 2013 from http://scholar.princeton.edu/wdobbie/files/Dobbie_Fryer_HCZ_II.pdf

Duckworth, A. L., (personal communication, May 1, 2013)

Duckworth, A. L., & Carlson, S. M. (in press). *Self-regulation and school success*. In B.W. Sokol, F.M.E. Grouzet, & U. Müller (Eds.), Self-regulation and autonomy: Social and developmental dimensions of human conduct. New York: Cambridge University Press.

Duckworth, A. L., Kirby, T. A., Gollwitzer, A., & Oettingen, G. (2013). From Fantasy to Action Mental Contrasting With Implementation Intentions (MCII) Improves Academic Performance in Children. *Social Psychological and Personality Science.* doi:10.1177/1948550613476307

Duckworth, A. L., Kirby, T. A., Tsukayama, E., Berstein, H., & Ericsson, K. A. (2011). Deliberate Practice Spells Success Why Grittier Competitors Triumph at the National Spelling Bee. *Social psychological and personality science, 2*, 174-181.

Duckworth, A. L., Peterson, C., Matthews, M. D., & Kelly, D. R. (2007). Grit: Perseverance and passion for long-term goals. *Journal of Personality and Social Psychology, 92*, 1087-1101.

Duckworth, A. L., Weir, D., Tsukayama, E., & Kwok, D. (2012). Who does well in life? Conscientious adults excel in both objective and subjective success. *Frontiers in Psychology, 3.*

Dweck, C.S. (1999). *Self-Theories: Their role in motivation, personality and development*. Philadelphia: Taylor and Francis/Psychology Press.

Dweck, C.S. (2006). *Mindset*. New York, NY: Random House.

Dweck, C.S., Walton, G.M., & Cohen, G. (2011). *Academic tenacity*. White paper prepared for the Gates Foundation. Seattle, WA.

Eisenberg, N., Duckworth, A. L., Spinrad, T. L., & Valiente, C. (2012). Conscientiousness: Origins in Childhood? *Developmental Psychology.* doi:10.1037/a0030977

Elmore, R. F., & Burney, D. (1997). *Investing in teacher learning: staff development and instructional improvement in Community School District #2,* New York City. Washington, DC: National Commission on Teaching and America's Future.

Elmore, R., & Burney, D. (1998). *Continuous improvement in Community District# 2, New York City*. Pittsburgh, PA: High Performance Learning Communities Project, Learning Research and Development Center, University of Pittsburgh.

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 noncognitive factors in shaping school performance: A critical literature review*. Chicago, IL: University of Chicago Consortium on Chicago School Research.

Fink, E., & Resnick, L. B. (2001). Developing principals as instructional leaders. *Phi Delta Kappan, 82*, 598–606.

Fullan, M. (2001). *The new meaning of educational change*. New York, NY: Teachers College Press.

Garcia, J. & Cohen, G. L. (2012). A social-psychological approach to educational intervention. In E. Shafir (Ed.), *Behavioral foundations of policy*, (pp. 329-350). Princeton, NJ: Princeton University Press.

Gomez, K., Lozano, M., Rodela, K., & Mancervice, N. (2012, November 8-11). Increasing access to mathematics through a literacy language lens. *Paper presented at the American Mathematical Association of Two-Year Colleges (AMATYC), Jacksonville, Florida.*

Hattie, J., & Timperley, H. (2007). *The power of feedback. Review of educational research, 77*, 81-112.

Haynes, T. L., Perry, R. P., Stupnisky, R. H., & Daniels, L. M. (2009). A review of attributional retraining treatments: Fostering engagement and persistence in vulnerable college students. In Smart, J. C. (Ed.), *Higher education: Handbook of theory and research* (pp. 227-272). New York, NY: Springer.

Hedengren, D., & Stratmann, T. (2012). The dog that didn't bark: What item non-response shows about cognitive and non-cognitive ability. Retrieved August 8, 2013 from http://papers.ssrn.com/sol3/papers.cfm?abstract_id=2194373

Hess, G. A. (1995). *Restructuring urban schools: A Chicago perspective*. New York, NY: Teachers College Press.

Hiebert, J., Gallimore, R., & Stigler, J. (2002). A knowledge base for the teaching profession: What would it look like, and how can we get one? *Educational Researcher, 31*, 3-15.

Holbrook, A. L., Krosnick, J. A., Carson, R. T., & Mitchell, R. C. (2000). Violating conversational conventions disrupts cognitive processing of attitude questions. *Journal of Experimental Social Psychology, 36*, 465-494.

Holbrook, A. L., Krosnick, J. A., Moore, D., & Tourangeau, R. (2007). Response Order Effects In Dichotomous Categorical Questions Presented Orally: The Impact of Question and Respondent Attributes. *Public Opinion Quarterly, 71*, 325-348

Hulleman, C. S., & Harackiewicz, J. M. (2009). Promoting interest and performance in high school science classes. *Science, 326*, 1410-1412.

Huston, L., & Sakkab, N. (2006). Connect and develop. *Harvard business review, 84*, 58-66.

Imai, M. (1986). *Kaizen (Ky'zen), the key to Japan's competitive success*. New York: Random House Business Division.

Institute for Healthcare Improvement. (2010). *90-Day Research and Development Process*. Retrieved from http://www.ihi.org/about/Documents/IHI90DayResearchandDevelopmentProcessAug10.pdf

Jacob, B., & Levitt, S. (2003). Rotten apples: An investigation of the prevalence and predictors of teacher cheating. *Quarterly Journal of Economics, 118*, 843–877.

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.

Karabenick, S. A. (2004). Perceived Achievement Goal Structure and College Student Help Seeking. *Journal of educational psychology, 96*, 569.

Knight, J. (2007). *Instructional coaching: A partnership approach to improving instruction*. Thousand Oaks, CA: Corwin Press.

Krosnick, J. A. (1991). Response strategies for coping with the cognitive demands of attitude measures in surveys. *Applied Cognitive Psychology, 5*, 213-236.

Krosnick, J. A. (1999). Survey research. *Annual Review of Psychology, 50*, 537-567.

Krosnick, J. A., & Alwin, D. F. (1987). An evaluation of a cognitive theory of response order effects in survey measurement. *Public Opinion Quarterly, 51*, 201-219.

Krosnick, J. A., & Fabrigar, L. R. (in press). *The handbook of questionnaire design*. New York: Oxford University Press.

Langley, G. J., Moen, R., Nolan, K. M., Nolan, T. W., Norman, C. L., & Provost, L. P. (2009). *The improvement guide: A practical approach to enhancing organizational performance* (2nd ed.). San Francisco, CA: Jossey-Bass.

Lewin, K. (1935*). A dynamic theory of personality: Selected papers*. New York, NY: McGraw-Hill.

Marat, D. (2005). Assessing mathematics self-efficacy of diverse students from secondary schools in Auckland: Implications for academic achievement. *Issues in Educational Research, 15*, 37-68.

Mazzocco M.M.M., & Kover, S.T. (2007). A longitudinal assessment of the development of executive functions and their association with math performance. *Child Neuropsychology, 13*, 18-45.

Merrow, J. (2013, April 11). Michelle Rhee's Reign of Error [Web log post]. Retrieved from http://takingnote.learningmatters.tv/?p=6232\

Mueller, C. M., & Dweck, C. S. (1998). Intelligence praise can undermine motivation and performance. *Journal of Personality and Social Psychology, 75*, 33–52.

Morris, A. K., & Hiebert, J. (2011). Creating Shared Instructional Products An Alternative Approach to Improving Teaching. *Educational Researcher, 40*, 5-14.

Narayan, S., & Krosnick, J. A. (1996). Education moderates some response effects in attitude measurement. *Public Opinion Quarterly, 60*, 58-88.

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.

Pfeffer, J., & Sutton, R. (2000). *The knowing-doing gap: How smart companies turn knowledge into action.* Cambridge, MI: Harvard Business School Press.

Presser, S., M. P. Couper, J. T. Lessler, E. Martin, J. Martin, J. M. Rothgeb, and E. Singer. 2004. Methods for testing and evaluating survey questions. *Public Opinion Quarterly 68*, 109–30.

Pyzdek, T., & Keller, P. A. (2003). *The Six Sigma handbook: a complete guide for green belts, black belts, and managers at all levels* (pp. 3-494). New York: McGraw-Hill.

Ramirez, G., & Beilock, S. (2011). Writing about testing worries boosts exam performance in the classroom. *Science, 331*, 211–213.

Rother, M. (2010). *Toyota kata: Managing people for improvement, adaptiveness, and superior results.* New York, NY: McGraw Hill.

Rutschow, E. Z., Cullinan, D., Welbeck, R. (2012). *Keeping students on course: An impact study of a student success course at Guilford Technical Community College*. New York, NY: MDRC. Retrieved May 13, 2013 from http://www.mdrc.org/sites/default/files/Keeping%20Students%20on%20Course%20Full%20Report.pdf

Rutschow, E. Z., Richburg-Hayes, L., Brock, T., Orr, G., Cerna, O., Cullinan, D., ... & Martin, K. (2011). *Turning the tide: Five years of achieving the dream in community colleges*. MDRC. Retrieved May 13, 2013 from http://www.mdrc.org/sites/default/files/full_593.pdf

Saris, W., Revilla, M., Krosnick, J. A., & Shaeffer, E. (2010). Comparing questions with agree/disagree response options to questions with item-specific response options. *Survey Research Methods, 4*, 61-79.

Schaeffer, E. M., Krosnick, J. A., Langer, G. E., & Merkle, D. M. (2005). Comparing the quality of data obtained by minimally balanced and fully balanced attitude questions. *Public Opinion Quarterly, 69*, 417-428.

Schoenfeld, A. H. (1988). When good teaching leads to bad results: The disasters of "well taught" mathematics courses. *Educational Psychologist, 23*, 145-166

Schuman, H., & Presser, S. (1981). *Questions and answers in attitude surveys: Experiments on question form, wording, and context*. New York, NY: Academic Press.

State of Georgia. (2011, July 5). *Deal releases findings of Atlanta school probe*. Retrieved July 29, 2013, from http://gov.georgia.gov/press-releases/2011-07-05/deal-releases-findings-atlanta-school-probe

Strother, S., Van Campen, J., and Grunow, A. (2013). *Community College Pathways: 2011-2012 Descriptive Report*. Retrieved from Carnegie Foundation for the Advancement of Teaching: http://www.carnegiefoundation.org/sites/default/files/CCP_Descriptive_Report_Year_1.pdf

Surowiecki, J. (2005). *The wisdom of crowds*. New York, NY: Random House.

The Design-Based Research Collective. (2003). *Design-based research: An emerging paradigm for educational inquiry*. Educational Researcher, 5-8.

Tourangeau, R., Couper, M., & Conrad, F. (2004). Spacing, Position, and Order: Interpretive Heuristics for Visual Features of Survey Questions. *Public Opinion Quarterly, 68*, 368-393.

Tourangeau, R., Rips, L.J., & Rasinski, K. (2000). *The Psychology of Survey Response*. Cambridge: Cambridge University Press.

Tuttle, C. C., Gill, B., Gleason, P., Knechtel, V., Nichols-Barrer, I., & Resch, A. (2013). *KIPP middle schools: Impacts on achievement and other outcomes.* Washington, DC: Mathematica Policy Research. Retrieved April, 26, 2013.

Tyack, D. B., & Cuban, L. (1995). *Tinkering toward utopia: A century of public school reform*. Cambridge, MA: Harvard University Press.

U.S. Department of Education, Institute of Education Sciences, National Center for Education Evaluation and Regional Assistance, Regional Educational Laboratory Southeast. (2011). *Measuring student engagement in upper elementary through high school: A description of 21 instruments.* (Issues & Answers Report, REL 2011–No. 098). Retrieved from http://ies.ed.gov/ncee/edlabs.

U.S. Department of Education, National Center for Education Statistics, Institute of Education Sciences. (2008). *Community colleges: special supplement to the condition of education 2008.* (NCES 2008-033). Retrieved from http://nces.ed.gov/pubs2008/2008033.pdf

Vansteenkiste, M., Lens, W., & Deci, E. L. (2006). Intrinsic versus extrinsic goal contents in self-determination theory: Another look at the quality of academic motivation. *Educational psychologist, 41*, 19-31.

Vaquero, L. M., & Cebrian, M. (2013). The rich club phenomenon in the classroom. *Scientific Reports, 3,* 1174.

Visher, M. G., Butcher, K. F., & Cerna, O. S. (2010). *Guiding developmental math students to campus services: An impact evaluation of the Beacon Program at South Texas College.* New York, NY: MDRC. Retrieved May 13, 2013 fromhttp://www.mdrc.org/sites/default/files/full_382.pdf

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.

Walton, G. M., & Cohen, G. L. (2011). A brief social-belonging intervention improves academic and health outcomes among minority students. *Science, 331*, 1447–1451.

Walton, G. M., & Spencer, S. J. (2009). Latent ability: Grades and test scores systematically underestimate the intellectual ability of negatively stereotyped students. *Psychological Science, 20*, 1132-1139.

Weiss, M. J., Visher, M. G., Wathington, H., Teres, J., & Schneider, E. (2010). *Learning communities for students in developmental reading: An impact study at Hillsborough Community College*. New York: MDRC. Retrieved May 13, 2013 from http://www.mdrc.org/sites/default/files/full_424.pdf

Wentzel, K. R., & Wigfield, A. (1998). Academic and social motivational influences on students' academic performance. *Educational Psychology Review, 10*, 155-175.

Yeager, D. S., & Dweck, C. S. (2012). Mindsets That Promote Resilience: When Students Believe That Personal Characteristics Can Be Developed. *Educational Psychologist, 47*, 302-314.

Yeager, D. S., & Krosnick, J. A. (2011). Does mentioning "some people" and "other people" in a survey question increase the accuracy of adolescents' self-reports? *Developmental Psychology, 47*, 1674-1679.

Yeager, D. S., & Krosnick, J. A. (2012). Does mentioning "some people" and "other people" in an opinion question improve measurement quality? *Public Opinion Quarterly, 76*, 131-141.

Yeager, D. S., & Walton, G. M. (2011). Social-psychological interventions in education: They're not magic. *Review of Educational Research, 81*, 267-301.

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."

Zeidenberg, M., & Scott, M. (2011). *The content of their coursework: Understanding course-taking patterns at community colleges by clustering student transcripts*. New York, NY: Community College Research Center, Teachers College, Columbia University.