Practical Measurement

David Yeager

University of Texas at Austin

Anthony Bryk

Hannah Hausman

Jane Muhich

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 Iris Lopez for her 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

It has become increasingly important to accelerate the capacity to learn in and through educational practice. The present methodological review explains why this type of learning requires a different type of measurement, one that is distinct from the measures commonly used for accountability or theory development. The article presents a theoretical framework for *practical measurement,* which is then illustrated through a case study. The case study shows how a practical theory and a set of practical measures were created to assess the causes of "productive persistence," the set of non-cognitive factors thought to affect community college developmental math student success. The article then explains how researchers and practitioners can use such measures for practical purposes, specifically, to *assess changes*, *predict* which students are at risk for problematic outcomes, and *set priorities* for improvement work. This article concludes by generalizing from this case study and proposing a future research and development agenda.

Keywords: Improvement Research, Measurement, Assessment, Research Methodology, Motivation.

Practical Measurement

Over a decade ago, educational researchers from the Consortium on Chicago School Research (CCSR) discovered that the single best predictor of eventual high school graduation was being “on-track” by the end of ninth grade—that is, that a student earned enough credits to be promoted to tenth grade and had no more than one semester F in a core course (Allensworth & Easton, 2005). Informed by this research, in 2003 Chicago Public Schools instituted an accountability system to reward or punish schools for their on-track rates. After an initial boost in 2003, there was no improvement over the subsequent four years. In 2002, only 59% of high school freshmen in Chicago were “on-track;” by 2007, that number was 57%. In 2009, however, CCSR began releasing brief, easily interpretable, monthly data reports regarding on-track rates to each school. This spurred remarkable improvement. Even without a comprehensive “program” for the city of Chicago, this practically-relevant data informed practitioners about whether their changes improved on-track rates—the leading indicator of eventual dropout. This coincided with a 25 percentage-point increase in on-track rates, to 82% in 2013, and graduation rates are following suit. Remarkably, the largest gains were found for the most disadvantaged students—African-American males—and improvements have been found in nearly every school in the city (Roderick, Kelley-Kemple, Johnson, & Beechum, 2014).

This story of reform is, at its heart, a measurement story. More generally, the field of educational measurement has evolved to address two needs: *accountability* and *theory development*. The former allows us to know, with precision, how well individual units (i.e., districts, schools, classrooms, or individual students) perform. The latter allows us to discern what might be causing under-performance in general and, as such, what might alleviate it. Practitioners who work with students on a daily basis, however, often require an additional type of information. They want to know how they can reliably improve learning in their classrooms for their particular students in the immediate term. Further, they need to accomplish this while managing the vast array of demands on time and attention posed by orchestrating classroom instruction. We argue in this paper that the activity of learning in and through practice to improve outcomes in the context of everyday instruction often requires a different type of measurement, which we term *practical measurement*. We illustrate why it is necessary, how to create it, and how to use it.

Because practical measures often have different uses than do measures for accountability or theory development, different procedures and criteria are often needed to create and evaluate such measures. An ideal practical measure is one that is rooted in causal theory, effectively supports the improvement of the local problem to be solved, and is as efficient as possible. As we will show, sometimes just a single question on a survey, informed by theory and customized for a given context, can be administered broadly and repeatedly in classrooms and inform continuous improvement efforts.

This paper, a narrative methodological review, proceeds as follows. First, we set the context for practical measurement by discussing new ways of thinking about educational change—what has been called *improvement research*. We explain what improvement research is and why it can assist practitioners to produce changes that are reliably effective. We then outline why practical measures are helpful for conducting improvement research and why measures created for accountability and theory development, although useful for different purposes, are often inadequate for improvement. Following this theoretical overview, we present a case study. We illustrate the creation and use of practical measures in the context of addressing failure rates in community college developmental math courses. We include original data as appropriate to support theoretical arguments (a great deal of methodological detail can be found in the online supplemental material[1]). We then conclude with 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, a 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 new organizational position, the instructional coach, was introduced into schools (Elmore & Burney, 1997, 1998; Fink & Resnick, 2001; Knight, 2007). What coaches actually needed to know and be able to do, as well as the requisite organizational conditions necessary for them to carry out their 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 those demands. The recent introduction of formal teacher evaluation protocols has greatly amplified principals' stress.

When policymakers were unsatisfied with the rate of school improvement, high-stakes accountability schemes were introduced, leading to many unintended consequences. The incidence of test-score cheating accelerated and select students were ignored, as accountability

[1] See http://tinyurl.com/practicalmeasurement

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.[2] 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 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 these ideas. Districts and states lacked the expertise and organizational capacity to support these changes at scale, and many policymakers ignored arguably the most important means for any reform to work—developing will for 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, 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, Fishman, Cheng, & Sabelli, 2011). Assuring efficacy

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

as this adaptive integration occurs, however, is rarely subject to systematic design-development 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 to enable an organization to learn from its own practices to continuously improve.[3] We know from numerous sectors, such as private industry and health care, that such inquiries in practices can, in some cases, transform promising change ideas into initiatives that achieve efficacy reliably at scale.

Improvement research taps a natural human bent to learn by doing. The notion of learning in practice has a long tradition that grew out of the works of 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 the resulting student work for evidence of positive change. Likewise, school faculty may examine together data on the effectiveness of current practices and share potential improvement ideas. Improvement science seeks to bring analytic discipline to design-development efforts and rigorous protocols to testing improvement ideas. In this way, "learning by doing" in individual clinical practice can culminate in robust, practical field knowledge (Hiebert, Gallimore, & Stigler, 2002).

[3] 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. (2009).

Several tenets form improvement research. The first is that, within complex organizations, *advancing quality must be integral in day-to-day work* (see, for example, a discussion of the Toyota Quality Management System in Rother, 2010). While this tenet may seem obvious, it actually challenges prevailing educational practice, whereby a select few conduct research, design interventions, and create policies, while many others do the actual work. Further, 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.[4] Thus, the second tenet is that 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. The third tenet is that improvement science embraces *a design-development ethic*. It places emphasis on learning quickly, at low cost, by systematically using evidence from practice to improve. 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). The cycle involves testing a change formulated as an inquiry plan (Plan), carrying out the plan (Do), analyzing results from the test (Study), and deciding what additional changes to make before conducting the next test (Act; see Langley et al., 2009).[5] These are done repeatedly and on an initially small and increasingly larger scale as learning accumulates.

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

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

[5] Also see http://www.ihi.org/knowledge/Pages/HowtoImprove/default.aspx

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.

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.*[6] Instead, the focal concern is whether positive outcomes can be made to occur reliably as

[6] To elaborate, intervention research is typically solution centered. Such research seeks 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 them. The focus of the research is on assembling robust change packages that can reliably produce improvements in targeted problems under diverse organizational conditions, with varied sub-groups of students, and for different practitioners. While intervention-focused studies seek to develop reliable causal inference about what occurred in some particular sample of conditions, improvement research aims to assure that measurable improvement in outcomes occur reliably under diverse conditions.

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.*

Measurement for Accountability, Theory Development and Practice Improvement

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 is discussed below and presented in Table 1.

Measurement for accountability. Global outcome data on problematic concerns—such as student drop-out rates or pass rates on standardized tests—are needed to understand the scope of the problem and to be able to 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 to identify exemplary or problematic individuals, e.g., districts, schools, teachers, as a means to take a specific action, such as giving a reward or imposing a sanction. Because this focus is on measuring individual cases, the psychometrics of accountability data place a high need for reliability at the individual level (Table 1).

Although measures for accountability assess outcomes of interest to policymakers and practitioners, they are limited in their usefulness for improvement research for several reasons. First, the data are typically collected after the end of some cycle—e.g., the end of the school year—which means that the people affected by a problematic set of procedures already have been harmed. Notably, the individuals who provide the data—e.g., failed students—will typically not benefit from their own data. Second, because accountability involves 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. This can make it difficult to use accountability measures to learn how to improve. 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 also are 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 include long, somewhat redundant question batteries that assess 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.

Further, there is a premium in academic research on novelty, which is often a prerequisite for publication. Consequently, academic measure development is often concerned with 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? (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, which are crucial functions 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 who attended West Point Military Academy and found that levels of grit actually *decreased* significantly over the four years at West Point (A. 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 part of their training). Instead, according to Duckworth, it is likely that they were comparing themselves to even "grittier" peers or role models and revising their assessment of themselves accordingly (for non-anecdotal examples, see Tuttle et al., 2013, or Dobbie & Fryer, 2013). This 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 assessing changes.

Measurement for improvement (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 often require *direct measurement of intermediary targets* (often "mediators" or "leading indicators") to evaluate ideas for improvement and to inform their continued refinement. In this regard, a person carrying out improvement research can ask questions such as: Is a student's motivation and grit actually improving in places where an instructional change has been introduced? and, Which students benefit most and under what set of circumstances?

Second, practical measurement often presses toward *greater specificity* in measurement. Educators need data closely linked to specific work processes and change ideas introduced in a particular context, rather than the more general theoretical concepts that may have motivated the practical changes. Third, increased validity can be gained from measures when *framed in a language targeted to the specific units focal for change,* and *contextualized around experiences common* to these individuals. Fourth, and most significant from a practical perspective, measures need to be *engineered to be embedded within the constraints of everyday school practice*. For example, a survey given to students routinely in classrooms would need to be brief, perhaps taking no more than 3 minutes. A description of these features can be found in Row 3 of Table 1 and in Table 2.

Uses of improvement measures. Practical measures serve several functions (Table 2). 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. Similar to a formative assessment, practical measures are not designed to assess final outcomes such as student failure rates. Instead, they are designed to provide information on changes in key processes thought to cause those final outcomes such as student motivation in a given week. Often the final outcomes will be traditional accountability measures. Practical measures are key intermediate signposts on the way toward these desired outcomes.

A second use for a practical measure is *predictive analytics*. This use will enable the researcher to answer questions in regard to which individuals or groups of individuals are at higher risk for problematic outcomes within a given setting.

A third use for practical measures is *priority setting*. When practitioners are engaged in improvement work, they must 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—that is, for less variability in performance—directs their attention toward weakening over time the predictive relationships discussed above.

Validity of practical measures as compared to measures for theory development or accountability. A key standard in educational and psychological testing (AERA et al., 1999), is that validity "refers to the use of a test for a particular purpose." Evaluating the validity of a measure, then, means that "sufficient evidence must be put forward to defend the use of the test for that purpose" (Sireci, 2007, p. 477; also see AERA et al., 1999; Cronbach, 1971; Kane, 1992; Messick, 1989). This standard applies to practical measurement just as it does to other forms of measurement. However, because practical measures serve somewhat different purposes than measures for theory development or accountability, different types of evidence are sometimes necessary in order to evaluate the validity of a practical measure.

For example, both practical measurement and measurement for theory development are interested in predicting student behavior. Yet while the purpose of a measure for theory development might be to reject the null hypothesis that that there is no predictive relation between a set of conceptual variables *in general,* the purpose of a practical measure might be to show that a given measure can be predictive of key outcomes when administered *in the context of a specific problem* to be solved. Practical measures impose additional constraints, many of them logistical, some of them empirical. It matters that the measure is not unbearably burdensome to administer with regularity, and that the practitioner who will use the data can clearly interpret its meaning, so as to inform their continuing improvement activities. All of these—contextual appropriateness, efficiency, potential for intuitive interpretation—affect the "validity" of a practical measure, because they affect its usefulness toward meeting the desired purpose of informing improvement (see AERA et al., 1999).

Case Study

We now wish to make these broad themes concrete by illustrating them with a case study. A tenet of improvement research is that it is problem-centered, and, thus, when illustrating this improvement work, we refer to efforts to address a critical educational problem: extremely-low success rates of developmental mathematics students in community college. This effort, carried out in 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 that work more broadly on changes to curriculum and instruction (for initial evidence of efficacy of the improvement network see Strother, VanCampen, & Grunow, 2013). We believe that the case study is useful for imagining how improvement research may be helpful for promoting student success with efficacy and reliability at scale.[7] Below, we provide some background on this improvement project, followed by information about how practical measures were created, validated, and used.

Improving Developmental Mathematics Outcomes in Community Colleges

The United States is unique in the world in providing a redemptive path to postsecondary education through the community college. Over 14 million students who are seeking opportunities for a productive career and better life are enrolled in community college. Community college students are more likely to be low income, the first in their family to attend college, an underrepresented minority, or underprepared for college (Bailey, Jenkins & Leinbach, 2005; Rutschow et al., 2011). Between 60% and 70% 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% 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, disproportionately low income or racial or ethnic minority,

[7] For detail on the data used in this case, see http://tinyurl.com/practicalmeasurement

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 experiencing a seemingly 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 oriented toward ambitious learning goals and culminating in the awarding of college math credit. Statistics and quantitative reasoning 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 and have served several thousand students.

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 that foster deeper learning. Traditional math curricula do relatively little to engage students' interest or demonstrate the relevance of mathematical concepts to everyday life (Carnevale & Desrochers, 2003; also see Hulleman &

Harackiewicz, 2009). So-called "non-cognitive" factors also play a role (see Dweck, Walton, & Cohen, 2011; Yeager & Walton, 2011). Many students have had negative prior math experiences, which have led to the belief, "I am not a math person" (Dweck, 2006). This belief can trigger anxiety and poor learning strategies when faced with difficult or confusing math problems (Beilock, Gunderson, Ramirez, & Levine, 2010; Blackwell, Trzesniewski, & Dweck, 2007; Haynes, Perry, Stupinsky & Daniels, 2009).). This is compounded for some students, e.g., women, African Americans, who are members of groups who have been stereotyped as "not good at math" (Cohen, Garcia, Purdie-Vaughns, Apfel, & Brzustoski, 2009; Walton & Spencer, 2009). Research also has found 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 or in writing (Gomez, Lozano, Rodela, & Mancervice, 2012; Schoenfeld, 1988).

To respond to these root causes, the Pathways sought to integrate a package of student activities and faculty actions that aimed to increase *productive persistence*, defined as the *tenacity* to persist, and the *learning strategies* to do so productively. The goal was to then carry out continuous improvement on that package of activities. However, the field had not agreed on the factors leading to productive persistence in community college developmental math, nor had the field agreed on what interventions faculty or course designers might implement to successfully promote it, or how to reliably measure their efficacy in the short term. Therefore, to begin to improve productive persistence, it was first crucial to agree on a common framework and then create 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, as noted at the outset. Thus, beginning improvement research requires the development of a “practical theory,” an easily interpretable conceptual framework of the system that affects student outcomes, that practitioners view as useful in guiding their work, and that remains anchored in the best available empirical research. This is especially important when scholars and practitioners have not agreed on how to define or categorize the multitude of ideas in a field, as in the case of “student success” or “non-cognitive factors.” In what follows, we explain and illustrate how we created a practical theory in this domain.

What is a practical theory for improvement? We begin by noting that a practical theory is not a *disciplinary* theory, as it does not seek to document novel features of human psychology or social or economic processes that shape the ways that 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, e.g., motivation among students of all levels of ability or of all ages. In the present case, a practical theory 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.

To reiterate, the virtue of a practical theory is not that it is new 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 his or her own lens. 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 a practical theory? In the case of productive persistence, we began with the assumption that much good work already had been done in both research and practice. We therefore started by determining 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" (Huston, & Sakkab, 2006; Institute for Healthcare Improvement, 2010) that scanned what was known in the field about the conceptual area. Ninety days is an arbitrary timeline, but it provides a discipline for driving toward a useful framework, given extant knowledge.

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 and keyword searches in the leading databases, e.g., Google Scholar, PsychInfo). Math faculty listed potential causes of productive persistence and then voted on those they believed were most crucial. Through these steps, we identified over 182 possible constructs. Such a list, however, is impractical and does not solve the problem of having no consensus about where to focus improvement efforts. Therefore, using the two filters presented below in which constructs were removed or aggregated, we reduced the list to five broad areas, with a handful of specific elements within each. See Figure 1.

Filter 1: Does the state of the science support the general importance of the concept? As a first pass to reducing a list of constructs in a practical theory, it can be helpful to rely on the published scientific record. We did this for our specific case study, asking: (a) Are there data, ideally from experiments or quasi-experiments, that support a *causal* interpretation of the concept?; (b) Is the concept distinct enough to yield practical, distinctive implications (not "self-efficacy for cognitive strategies" vs. "self-efficacy for self-regulated learning"), and is it theoretically precise enough to be useful (i.e., not just "feeling connected" to the classroom)?; and (c) Is the concept an underlying cause, or is it better viewed as a downstream consequence of some directly changeable concept that causally precedes it? (i.e., although self-efficacy is highly predictive of behavior and important, the practical theory focused on the causal antecedents of self-efficacy, such as a fixed vs. growth mindset, Dweck, 1999). Answering these questions 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*?** A second filter to apply when narrowing a list of concepts for a practical theory is more extensive and fine-grained. This is used to customize the concepts more for the specific problem of practice to be solved—in the present case community college developmental math. The second filter included the following questions: (a) Is the concept likely to be amenable to change via the *systems of influence in place in the improvement setting,* e.g., either by a faculty member's behaviors or by the structure of the course? For instance, the time demands of being responsible for the care of children may contribute to lower 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 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 (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 that 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 can be created. The model can then be "tested" and refined by using focus groups and conversations with faculty, researchers, college counselors, and students. In our case, in these "testing" conversations practitioners expressed their opinion in regard to (a) whether they felt that the framework captured important influences on developmental math achievement; and (b) whether the concepts that comprise the framework were described in a way that made them understandable and conceptually distinct. We did this, and it 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.[8] The product of this effort, still a work in progress, is depicted in Figure 1.

Formulating a Practical Measure

A practical theory allows researchers to work with practitioners on a narrower, agreed-upon set of high-leverage factors thought to influence an outcome of interest. Use of the

[8] The meeting in which the practical theory was discussedvetted 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.

practical framework, however, requires implementing practical measures of the elements described in the framework. In the present case, after we identified and refined the five conceptual areas relevant to productive persistence (Figure 1), we then created a set of items to assess each. Because many of the ideas in the concept map had come, at least in part, from the academic literature, there were measures available for consideration. A comprehensive scan of research in the field located roughly 900 different potential survey measures.

By and large, however, available measures failed the test of practicality. Many items were redundant, theoretically-diffuse, double-barreled questions that used vocabulary that would be confusing for respondents who were English learners or who had 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, Fernandez, Jacobson, & Smither-Wulsin, 2012; for a similar review, see U.S. Department of Education, 2011), located 196 survey instruments, each with a number of individual items. Our team of coders reviewed each survey instrument but could not locate any evidence of predictive validity with objective outcomes (correlations with test scores or official grades) for 94% of the measures, even though statistics such as internal consistency reliability or common factor loadings were available for nearly all. Administration in community college populations was even rarer; our team could find only one paper that measured the concepts identified in our practical theory and showed a relationship to objective course performance metrics among developmental mathematics students. Of course, many of these self-report 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 spend no more than 3 minutes to answer survey questions. This created a target of approximately 25 survey items that could be used to assess the major constructs presented in Figure 1 and serve each of the purposes of practical measurement, i.e., assessing changes, predictive analytics, and setting priorities.). Therefore, our team took the list of 900 individual survey items uncovered in our review of the literature and reduced them to roughly 25 items that, in field tests with community college students, took an average of 3 minutes to answer. The steps for doing this are outlined below.

Step 1: Guided by theory. The process of narrowing and refining the overly-long list of practical measures can begin by looking to the experimental literature to learn what effectively promotes the desired outcome—in the present case, tenacity and the use of effective learning strategies, the hallmarks of productive persistence. One can then select or rewrite items so that they tap more precisely into causal theory.

For instance, while an enormous amount of important correlational research has focused on the impact of social connections for motivation (Wentzel & Wigfield, 1998), only a few experimental studies in social psychology focus more precisely on the concept of "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 he or she belongs in a class or a college, it can be difficult for that person to fully commit to the behaviors that may be necessary to succeed, such as joining a study group or asking professors for help. Of significance to practical measurement, it has been demonstrated that an experimental intervention that alleviates 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 presented 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 finding 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 on promoting relevance (Hulleman & Harackiewicz, 2009), supporting autonomy (Vansteenkiste, Lens, & Deci,, 2006), teaching 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 (Jamieson, Mendes, Blackstock, & Schmader, 2010; Ramirez & Beilock, 2011), and others. We then found or rewrote items that were face valid (to community college faculty and students) and precisely related to theoretical concepts that were malleable and likely to have leverage for affecting outcomes.

Step 2: Optimal item design. In addition to selecting theoretically-precise items, when creating practical measures it can be helpful to revise the wording of the items according to optimal survey design principles so as to maximize the amount of information that could be obtained from very few questions (see Krosnick, 1999; Schumann & Presser, 1981). In fact, there is a large amount of experimental literature in cognitive and social psychology that reports on practical measures in different settings, e.g., 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 also must be brief and clear. Notably, verbal administration can exaggerate the differences in measurement accuracy among low-education respondents (Holbrook, Krosnick, Moore, & Tourangeau, 2007; Krosnick & Alwin, 1987). A large number of national experiments have determined how to maximize accuracy and minimize response time for low-education sub-groups in particular (Narayan & Krosnick, 1996; see Krosnick, 1999). Such findings are relevant for administration to students who take developmental math in community college because they are, by definition, low-education respondents.

A number of lessons from the public opinion questionnaire design literature are relevant. One strong recommendation is, whenever possible, to 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, research has shown 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 (Krosnick, 1991), who, as noted, are the targets of our measures. Therefore, unless we 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),[9] we wrote "construct-specific" items.

What is a "construct-specific" item? A question that concerns math and statistics anxiety 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 a construct-specific format as, "How anxious would you feel taking a math or statistics test?" (response options: 1 = *not at all anxious*; 5 = *extremely anxious*). In fact, we tested these two response formats. We conducted a large-sample ($N > 1{,}000$) experiment for which we randomly assigned developmental math students to answer a series of items that assessed anxiety by an either agree/disagree or construct-specific format, 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, approximately 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, *r*s = .06 and .09, *n.s.*, respectively (and these correlations differed from one another, interaction effect *p*s < .05), which demonstrates significantly lower validity for agree/disagree items compared to construct-specific items in that case.

[9] 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 (see also. Saris, Revilla, Krosnick, & Shaeffer, 2010; Schumann & Presser, 1981).

In writing effective practical measurement questions it can be helpful to employ a number of additional “best practices” for reducing response errors among low-education respondents. These included our 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, as laboratory experiments show that response quality is greater on 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 questions with, “Some people think . . . but other people think . . . ” (Yeager & Krosnick, 2011; 2012). In our case study, we followed all of these recommendations.

Step 3: Contextualizing and pre-testing. After an initial period of item writing, the next step in creating a practical measure is to customize them so that they feel contextually-relevant to the target group. Hence, in our case study we next customized the survey items to the perspectives of community college practitioners and students. Following best practices, we also conducted cognitive pretests (Presser et al., 2004) with current developmental math students to bring to light 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, groups that would be especially likely to underperform in developmental math and, therefore, groups from which, ideally, the practical measures would help us learn the most about how to help. This led to our rewriting of a number of items and to confirmation that many survey items were successfully eliciting the type of thinking that 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 programs, however, not all of these items proved to be predictive of student outcomes, on either an individual or 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 four items). In such cases, one item was written for each distinct thought or experience and then combined into the higher-level construct.). Altogether, 15 survey items were used to measure the following five constructs (please see the online supplement for exact wording and response options: http://tinyurl.com/practicalmeasurement).

It is important to note that, while these items are 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 each item measures aspects of the practical theory, as presented in Figure 1, some measures that we created did not show meaningful validity correlations. Thus, further development is needed to more fully measure all of the concepts in the practical theory. Nevertheless, the resulting practical measure illustrates the uses of practical measures, as we discuss 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, which is a website that hosts 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 (for exact response rates, see the online supplement: http://tinyurl.com/practicalmeasurement).

Examples of Practical Measurement to Improve

Three uses of practical measures are (a) assessment of changes, (b) predictive analytics and (c) priority setting. We illustrate each of these below in the context of our case study and then summarize key differences in Table 3. Primary data come from a sample of 1,391 students in two developmental math pathways (the Statway® and the Quantway®) at over 30 colleges; 78% placed two or more levels below college-level math, and two-thirds were racial or ethnic minorities (Hispanic or African American). The findings were then replicated in the same math pathways in the subsequent year, with a new sample of *N* = 1,217 students (see online supplement for greater detail). Hence the findings in this case study are reproducible.

1: Assessment of change. One use for practical measures is to assess whether the changes implemented were, in fact, improvements, at least in terms of the concepts outlined in the practical theory. Here we explain how our practical measures were used for this purpose in the context of efforts to promote productive persistence.

Evaluating a "Starting Strong" package. First, we assessed changes in productive persistence factors in response to activities in the beginning of the course. Both practitioner accounts and empirical studies show that the first few weeks of a 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, they may withhold the effort that is required to have success in the long term, which starts a negative recursive cycle that ends in either course withdrawal or failure (Cook, Purdie-Vaughns, Garcia, & Cohen, 2012; Vaquero & Cebrian, 2013). 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 as well as strikingly predictive of student learning over time (Vaquero & Cebrian, 2013). The productive persistence conceptual framework posits that, if faculty successfully create a classroom climate that helps students to see their academic futures as more hopeful and that facilitate the development of 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.[10]

In light of these possibilities, the productive persistence activities involved 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), including reducing anxiety, increasing interest in the course, and forming supportive student social networks. The "Starting Strong" package also included a brief, one-time "growth mindset" reading and writing activity that had been shown in past experimental research to increase overall math grades among community college students (Blackwell et al., 2007; Yeager & Dweck, 2012; Yeager, Paunesku, Walton, & Dweck, 2013). There were also classroom activities for forming small groups, and encouraging peers to exert positive peer pressure to increase class attendance.

To determine whether the practical measures were effective at assessing changes, we examined the productive persistence survey on the first day of class and after three weeks of instruction. Evidence for the efficacy of the Productive Persistence "Starting Strong" package, presented in Figure 2, 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

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

exposure to Statway®. As instruction began, students' interest in math increased, their belief about whether math ability is a fixed quantity decreased, and math anxiety and their uncertainty about belonging decreased as well. Although caution is warranted in making causal interpretations from change data, the data were given a chance to "object" and they did not. On average, they are consistent with the prediction that the theory-based changes introduced into classrooms could improve student persistence through the early weeks of instruction. From an improvement perspective, these results suggest that continued efforts in this area are warranted. In addition, we also learned that these changes did not occur in every college and for every sub-group of students. The latter results, in conjunction with predictive validity findings, informed subsequent improvement priority setting (see below).

2: Predictive analytics.

At-riskness index. Another use for practical measures is to assess initial levels of risk of showing a problematic outcome (also see Roderick et al., 2014). In our case, we asked whether it was possible to create 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 has the potential to 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: (a) skills and habits for succeeding in college; (b) students' belief that they are capable of learning math; and (3) students' feeling socially tied to peers, faculty, and course of study. Data on the perceived value of the course were not included in the at-riskness indicator because the course content was already designed to engage disinterested students and explain the relevance of math and statistics. The measures of faculty's mindsets and skills also were 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.

Our analyses produced empirically derived cut points that signaled problematic versus non-problematic responses on five different risk factors (anxiety, mindsets about academic ability, social ties, stereotype threat, and "grit") for the three concepts listed above. We then summed the number of at-risk factors to form an overall at-riskness score that ranges from 0 to 5. The systematic procedure for doing this is presented in great detail online (http://tinyurl.com/practicalmeasurement).

As illustrated in Figure 3, productive persistence risk level on the first day of the course showed a striking relationship to course outcomes (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. Indicative of the robustness of these findings, they were replicated in both the Statway® and the Quantway® colleges. Further, 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/ethnicity 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 vs. a much greater chance of successfully completing the course (Figure 3).

Real-time student engagement data. The analyses presented above show that it is possible to identify *students* with higher levels of risk for not productively persisting. We also wondered, however, whether it was 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 average course failures, interventions might be developed to help instructors to keep students engaged.

As a first step toward doing this, the Carnegie Foundation instituted very brief (3- to 5-question) "pulse check" surveys in the online instructional system, which, as noted, is the website that 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, they were redirected to a single-page, optional survey that consisted of three to five items. Students were asked their views about the specific lessons just studied (whether there were language issues, whether it was interesting and relevant). Most crucially for the present purposes, they were asked one key summary question "Overall, how do you feel about the Statway course right now?" (1 = *extremely negative*, 2 = *mostly negative*, 3 = *mostly positive*, 4 = *extremely positive*). All student responses within a class were averaged on this summary question to produce a mean score for each class for each day. As shown in Figure 4, nearly all classrooms began with high levels of enthusiasm. This cooled over time toward a 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. In contrast, less successful classes showed a continued downward decline, with students' becoming more negative toward the course as the term progressed.

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 (Figure 4). If future analyses replicated these trends across contexts, it would be easy to see how this practical measure could constitute an effective early warning system for targeting classroom-level improvement efforts, such as professional development for 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 been successfully addressed. For instance, we found that one survey item, which assessed belonging uncertainty, administered in the fourth week of the course (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; regression table presented in the online supplement). Further, 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 (achieving a grade of B- or better for the first semester; Figure 5).

These findings have led directly to priorities for improvement efforts to address belonging uncertainty. These data were a signal to faculty that, in their classes, belonging uncertainty was not being sufficiently addressed but mattered a great deal for their students. 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 context an experimental social-psychological intervention that has a 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 PDSA cycles

(Deming, 1986; Imai, 1986; Morris & Hiebert, 2011; Pyzdek & Keller, 2009) on new routines to create 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, e.g., routines for emailing absent students, improved routines for creating and maintaining collaborative small-groups. Faculty track practical measures of behaviors, such as attendance, and periodically administer the survey items that assess 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 withdraw or fail. 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, which, ideally, will lead to accelerated and reliable improvements in student outcomes at scale.

Other Cases of Practical Measurement

The present case study—focused on productive persistence within developmental math—is but one example of practical measurement. Other efforts are informative. Within education, organizations such as the Chicago Consortium for School Research has found that a small set of survey items assessing trust in school can be among the best indicators of a school's likelihood of improving over time (see Bryk, Sebring, Allensworth, Luppescu, & Easton, 2010). These have been embedded in regularly-administered surveys, directly informing school improvement efforts and allowing for assessments of change over time. This is in addition to the 9th grade "on-track" indicator, which has informed improvement of high school graduation rates (Roderick et al., 2014). Although these efforts are not carried out in the context of a full networked improvement community, they illustrate the usefulness of practical measurement for informing organizational change.

Outside of education, marketing research in particular and opinion research more generally (e.g., Krosnick, 1999) has developed increasingly informative methods for practically measuring individuals' consumption behavior or product preferences. For instance, there is a large industry devoted to improving and understanding a single-item self-report: the so-called "net promoter" value, which asks consumers to state on an 11-point scale how likely they would be to recommend a given product to a friend (Reichheld, 2003). In analyzing the net promoter score, a marketing researcher creates a percentage by dividing the percent of people giving strong, positive responses by the percent of people giving more negative responses. Because the purpose of this type of improvement research is to track an aggregate of people's attitudes toward a product, and not reliably assess an individual person's attitudes, then it can be possible to have informative, practical data with very brief, single-item measures (Reichheld, 2003). This use closely parallels our "pulse check" survey to track classroom attitudes depicted in Figure 4.

Perhaps most influential has been research by the Institute for Healthcare Improvement (IHI), which has developed a systematic framework for creating and using practical measures to inform improvement. That is, instead of administering long and redundant survey items in hospitals, an improvement effort can be informed by "passive" data already collected by the hospital—such as patient infection rates. Teams of improvement researchers can track these data and drive down problematic outcomes through repeated PDSA cycles. This IHI framework has been generalized to many other areas, such as reducing energy use, reducing contamination in shipping drums, or improving service in a dental office (see Langley et al., 2009). IHI efforts provide a directly informative analogy to burgeoning improvement efforts in education.

Using Practical Measures: Who and in What Context?

Who are the individuals who will actually use practical measures? Both within education and outside of education, most improvement occurs in an informal or formal improvement team (Langley et al., 2009). While there is no single way for this team to be organized, it usually involves the following expertise: (a) deep theoretical knowledge of the concepts outlined in the practical framework (e.g., Figure 1); (b) practical and theoretical knowledge of improvement methods, such as PDSA cycles and practical measurement; and (c) practical expertise in the problem to be solved and the context in which it will be addressed (also see Yeager & Walton, 2011). Often, these different types of expertise reside within disciplinary researchers, improvement specialists, and improvement-oriented practitioners, respectively, all of whom work toward improving a common outcome (for greater detail on how improvement teams are often organized, see Langley et al., 2009).

Future Directions for Research on Practical Measurement

Behavioral assessments. In this paper, we have presented the development and use of predominantly self-reported practical measures of productive persistence. We have done this because a great deal of research has supported the assertion that, if you ask people sensible questions to which they know the answer, under circumstances in which they feel able to report their true opinions, you can gather highly predictive data from even brief sets of questions (Krosnick, 1999).

In many cases, however, it also would be desirable to develop behavioral indicators of the concepts outlined in a practical theory to supplement these self-reports. This is true, in part, due to reference bias (Biernat, 2003), which is the tendency for a self-reported measure to rely on the subjective frame of reference of the respondent (recall the example of “grit” at West Point

Military Academy; for other instances where reference bias may have occurred in the assessment of non-cognitive factors, see Dobbie & Fryer, 2013; Tuttle et al., 2013).

To bypass some of the potential limitations of self-reports, in some cases, it is desirable to design novel behavioral measures. Returning to productive persistence as an example, one could analyze whether students review their work 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, an individual who conducts improvement research could embed opportunities for student choice in the level of difficulty of tasks and then determine the percentage of students who chose hard tasks, from which they could learn a lot, as opposed to easy tasks, from which they could earn a high score (for examples of such measures, see Mueller & Dweck, 1998).

In addition, 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, whether students ask for help, or whether students self-remediate when confused, to indicate non-cognitive concepts, as shown in Figure 1 (see our behavioral measure of "grit" in the at-riskness index; see also Hedengren & Stratmann, 2012). More generally, so-called "passive" 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 logging into an online platform or clicking through problems versus honestly attempting them, might be a rich source of data that can inform improvement efforts in education.

Psychometric issues. 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 as a means to avoid both false negatives and false positives that might unfairly affect a teacher, school, or district. Such uses emphasize 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 for either accountability or theory (e.g., Cronbach's α, model fit in a confirmatory factor analysis).

Because, as noted earlier, the purposes of 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 a practical measure. An ideal 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 and only modestly correlated. Items that measure different constructs also would ideally have no loading on a common factor because there would be no redundant measures or clusters of items in the battery; no respondent time would be wasted. In contrast, predictive validity, or whether the measure predicts long-term outcomes of interest, is at a premium, as is the potential for the measure to be sensitive to even small changes in instruction or classroom culture, e.g., those on a weekly basis. Indeed, Cronbach (1961) himself stated, “If predictive validity is satisfactory, low reliability does not discourage us from using the test” (p. 128).[11]

Even test-retest reliability is not certain to inform the likelihood that a measure will be valid. In the absence of effective intervention, a practical measure should have strong test-retest reliability and predict long-term outcomes. Yet in the presence of an effective intervention, a practical measure’s predictive relation and test-retest reliability should be driven to zero because the risk factor would have been successfully addressed.

One helpful analogy comes from a practical measure in use in hospitals all over the world: the Apgar score (Apgar, 1953). This score involves easy-to-observe features of 1-minute-old newborn infants, and it can be calculated quickly by any competent practitioner. Low scores on the Apgar can signal a “failure to thrive.” Fortunately, however, this risk is rarely translated into reality because doctors and nurses are usually able to immediately intervene to give the newborn the treatment it needs. Low 1-minute Apgar scores do not strongly predict deficits later in development (Casey, McIntire, & Leveno, 2001), but this does not mean that the scores are not valid. Instead, this illustrates how, in a well-functioning system, an effective “at-riskness” practical measure directly informs intervention, and so its predictive validity is driven to zero. Interestingly, if a newborn still has a low Apgar score after 5 minutes, then the newborn has a high risk of long-term health problems or even death (Casey et al., 2001). Thus, sometimes-practical measures can be re-administered quickly and can signal whether an intervention was effective at preventing negative outcomes.

[11] Interestingly, Bergkvist and Rossiter (2007) showed *no* improvement in predictive validity for a multiple-item measure as compared to single-item measures, which raises questions about whether, even in measurement for theory development, long and redundant batteries are necessary to produce validity (also see a striking meta-analysis by McCrae, Kurtz, Yamagata, & Terracciano, 2011).

Altogether, a different or revised set of psychometric rules might be needed for improvement researchers or reviewers of manuscripts and grant proposals that include practical measures and improvement research. This may be an especially high priority given that, in recent years, there has been an increase in calls for grant proposals related to improvement research, and experts require guidance for evaluating the strength of these proposals.[12] Clearly, much more theoretical work, simulation work, and field applications are required to do this sufficiently.

Conclusion

We have argued that educators need to be able to assess whether the instruction that they deliver in a classroom does, in fact, lead to the changes that they hope for, in real time, well before students become academic casualties. Although measurement for accountability is important for signaling the presence of 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 requires deeper information about system processes and where undesirable outcomes stem from as well as targeting subsequent improvement based on this knowledge. Seeking to remediate the problem at the end of the line is not an effective solution (Rother, 2010).

Educators need both more frequent data and different kinds of information than they normally have; they need measures that can help them improve their actual practices. Yet the measures that the field has developed for theory development are often impractical. We look forward to future research on methods to create and embed practical measures in networks of

[12] See http://ies.ed.gov/funding/pdf/2014_84305H.pdf

researchers and practitioners engaged in improvement research. We believe that this can play a substantial role in the quality improvement of educational processes at scale.

References

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1999). *Standards for educational and psychological testing.* Washington, DC: American Educational Research Association.

Apgar, V. (1953). A proposal for a new method of evaluation of the newborn infant. *Current Research in Anesthesia & Analgesia, 32,* 260–267.

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 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. doi:10.1016/j.econedurev.2009.09.002

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 of the United States of America, 107*, 1860–1863. doi:10.1073/pnas.0910967107

Bergkvist, L., & Rossiter, J. R. (2007). The predictive validity of multiple-item versus single-item measures of the same constructs. *Journal of Marketing Research, 44,* 175-184.

Berwick, D. M. (2008). The science of improvement. *The Journal of the American Medical Association, 299*, 1182-1184. doi:10.1001/jama.299.10.1182

Biernat, M. (2003). Toward a broader view of social stereotyping. *American Psychologist, 58*, 1019-1027. doi:10.1037/0003-066X.58.12.1019

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. doi:10.1111/j.1467-8624.2007.00995.x

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. S., Sebring, P. B., Allensworth, E., Luppescu, S., & Easton, J. Q. (2010). *Organizing schools for improvement: Lessons from Chicago.* Chicago, IL: University of Chicago Press.

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

Carnevale, A. P., & Desrochers, D. M. (2003). *Standards for what? The economic roots of K–16 reform.* Princeton, NJ: Communication and Public Affairs, Office of Assessment, Equity, and Careers, Educational Testing Service.

Casey, B. M., McIntire, D. D., & Leveno, K. J. (2001). The continuing value of the Apgar score for the assessment of newborn infants. *New England Journal of Medicine, 344,* 467–471.

Chang, L., & Krosnick, J. A. (2010). Comparing oral interviewing with self-administered computerized questionnaires: An experiment. *Public Opinion Quarterly, 74,* 154–167. doi:10.1093/poq/nfp090

Clark, C. A., Pritchard, V. E., & Woodward, L. J. (2010). Preschool executive functioning abilities predict early mathematics achievement. *Developmental Psychology, 46,* 1176–1191. doi:10.1037/a0019672

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. doi:10.1126/science.1170769

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. doi:10.1037/a0026312

Cronbach, L. J. (1961). *Essentials of psychological testing.* New York, NY: Harper and Row.

Cronbach, L. J. (1971). Test validation. In R. L. Thorndike (Ed.), *Educational measurement* (2nd ed., pp. 443–507). Washington, DC: American Council on Education.

Deming, W. E. (1986). *Out of the crisis: Quality, productivity, and competitive position.* Cambridge, MA: Cambridge University Press.

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 from http://scholar.princeton.edu/wdobbie/files/Dobbie_Fryer_HCZ_II.pdf

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, NY: Cambridge University Press.

Duckworth, A. L., Kirby, T. A., Gollwitzer, A., & Oettingen, G. (in press). 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. doi:10.1177/1948550610385872

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. doi:10.1037/0022-3514.92.6.1087

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*(356), 1-8. doi:10.3389/fpsyg.2012.00356

Dweck, C. S. (1999). *Self-theories: Their role in motivation, personality and development*. Philadelphia, PA: 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. (in press). 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 & America's Future.

Elmore, R. F., & 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.

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). *Increasing access to mathematics through a literacy language lens.* Paper presented at the American Mathematical Association of Two-Year Colleges (AMATYC), Jacksonville, FL.

Hattie, J., & Timperley, H. (2007). The power of feedback. *Review of Educational Research, 77*, 81–112. doi:10.3102/003465430298487

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. doi:10.1007/978-1-4020-9628-0_6

Hedengren, D., & Stratmann, T. (2012). The dog that didn't bark: What item non-response shows about cognitive and non-cognitive ability. doi:10.2139/ssrn.2194373

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

Hiebert, J., Gallimore, R., & Stigler, J. W. (2002). A knowledge base for the teaching profession: What would it look like, and how can we get one? *Educational Researcher, 31*, 3–15. doi:10.3102/0013189X031005003

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. doi:10.1006/jesp.1999.1411

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. doi:10.1093/poq/nfm024

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

Huston, L., & Sakkab, N. (2006). Connect and develop: Inside Procter & Gamble's new model for innovation. *Harvard Business Review, 84,* 58–66.

Imai, M. (1986). *Kaizen (Ky'zen): The key to Japan's competitive success*. New York, NY: McGraw-Hill.

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. *The Quarterly Journal of Economics, 118*, 843–877. doi:10.1162/00335530360698441

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. doi:10.1016/j.jesp.2009.08.015

Kane, M. T. (1992). An argument-based approach to validity. *Psychological Bulletin, 112,* 527–535.

Karabenick, S. A. (2004). Perceived achievement goal structure and college student help seeking. *Journal of Educational Psychology, 96*, 569–581. doi:10.1037/0022-0663.96.3.569

Kim, N., Krosnick, J., & Casasanto, D. (2012). *Moderators of candidate name order effects in elections: An experiment.* Unpublished manuscript. Stanford, CA: Stanford University Press. An unpublished manuscript would not have a publisher. Please clarify.

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

Krosnick, J. A. (1999). Survey research. *Annual Review of Psychology, 50,* 537–567. doi:10.1146/annurev.psych.50.1.537

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. doi:10.1086/269029

Krosnick, J. A., & Fabrigar, L. R. (in press). *The handbook of questionnaire design*. New York, NY: 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 executive function skills and their association with math performance. *Child Neuropsychology, 13*, 18–45. doi:10.1080/09297040600611346.

McCaffrey, D. F. (2012, October 15). Do value-added methods level the playing field for teachers? [Newsgroup]. Retrieved from http://www.carnegieknowledgenetwork.org/briefs/value-added/level-playing-field/

McCrae, R. R., Kurtz, J. E., Yamagata, S., & Terracciano, A. (2011). Internal consistency, retest reliability, and their implications for personality scale validity. *Personality and Social Psychology Review, 15*, 28–50. doi:10.1177/1088868310366253

Messick, S. (1989). Meaning and values in test validation: The science and ethics of assessment. *Educational Researcher, 18,* 5–11.

Morris, A. K., & Hiebert, J. (2011). Creating shared instructional products: An alternative approach to improving teaching. *Educational Researcher, 40*, 5–14. doi:10.3102/0013189X10393501

Mueller, C. M., & Dweck, C. S. (1998). Praise for intelligence can undermine motivation and performance. *Journal of Personality and Social Psychology, 75*, 33–52. doi:10.1037/0022-3514.75.1.33

Narayan, S., & Krosnick, J. A. (1996). Education moderates some response effects in attitude measurement. *Public Opinion Quarterly, 60,* 58–88. doi:10.1086/297739

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. doi:10.3102/0013189X11421826

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

Presser, S., Couper, M. P., Lessler, J. T., Martin, E., Martin, J., Rothgeb, J. M., & Singer, E. (2004). Methods for testing and evaluating survey questions. *Public Opinion Quarterly, 68*, 109–130. doi:10.1093/poq/nfh008

Pyzdek, T., & Keller, P. A. (2009). *The Six Sigma handbook: A complete guide for green belts, black belts, and managers at all levels* (3rd ed., pp. 3-494). New York, NY: McGraw-Hill.

Ramirez, G., & Beilock, S. L. (2011). Writing about testing worries boosts exam performance in the classroom. *Science, 331*, 211–213. doi:10.1126/science.1199427

Reichheld, F. F. (2003). The one number you need to grow. *Harvard Business Review, 81*, 46–54.

Roderick, M. Kelley-Kemple, T., Johnson, D.W., & Beechum, N. O. (2014). *Preventable failure: Improvements in long-term outcomes when high schools focused on the ninth* *grade year*. Chicago, IL: The University of Chicago Consortium on Chicago School Research.

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

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.* New York, NY: MDRC. Retrieved from Casey Foundation website: http://imap.caseyfoundation.net/~/media/Pubs/Topics/Economic%20Security/Family%20Economic%20Supports/TurningtheTideFiveYearsofAchievingtheDreaminCommunityColleges/MRDCcollege.pdf

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

Schoenfeld, A. H. (1988). When good teaching leads to bad results: The disasters of "well taught" mathematics courses. *Educational Psychologist, 23*, 145-166. doi:10.1207/s15326985ep2302_5

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

Shaeffer, 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. doi:10.1093/poq/nfi028

Sireci, S. G. (2007). On validity theory and test validation. *Educational Researcher, 36,* 477–481.

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

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

The Design-Based Research Collective. (2003). Design-based research: An emerging paradigm for educational inquiry. *Educational Researcher, 32,* 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. doi:10.1093/poq/nfh035

Tourangeau, R., Rips, L. J., & Rasinski, K. (2000). *The psychology of survey response*. Cambridge, MA: Cambridge University Press. doi:10.1017/CBO9780511819322

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., & 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. doi:10.1207/s15326985ep4101_4

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

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

Walton, G. M., & Cohen, G. L. (2011). A brief social-belonging intervention improves academic and health outcomes of minority students. *Science, 331,* 1447–1451. doi:10.1126/science.1198364

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. doi:10.1111/j.1467-9280.2009.02417.x

Wentzel, K. R., & Wigfield, A. (1998). Academic and social motivational influences on student's academic performance. *Educational Psychology Review, 10,* 155–175. doi:10.1023/A:1022137619834

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. doi:10.1037/a0025440

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. doi:10.1093/poq/nfr066

Yeager, D.S. & Dweck, C.S. (2012). Mindsets that promote resilience: When students believe that personal characteristics can be developed. *Educational Psychologist, 47,* 1-13.

Yeager, D.S. & Walton, G. (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.* 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.

Table 1 *Types of Measurement*

Measurement Focus	Typical Use	Sample Research Question	Common Features	Implications for Psychometrics	Limitations for Improving Practice
Accountability	Identifying exemplary or problematic individual teachers, schools, or districts.	“Which schools should we put on probation?”	Summative, global performance measures, typically collected once a year, often toward the end of the academic year.	Extremely high reliability at the level at which you are rewarding or punishing.	Data are typically reported after school year has concluded. Students providing data do not directly benefit because data refer to last year’s instruction/teacher/ curriculum. Causes of differences are opaque and not tied to specific practices.
Theory development and testing	Test a theory regarding the relations among two or more conceptual variables.	“Does low self-efficacy predict less learning?”	Goal is to detect stable individual differences among students, teachers, or schools on the constructs of interest. Administer long, somewhat redundant surveys assessing multiple small variations on the some concept. Typically used to maximize estimated relations between latent variables.	High internal consistency reliability and construct validity as assessed via factor analyses. Goal of minimizing error variance in construct measure is key to goal of maximizing estimated relations between latent variables.	Impractical to administer as a part of standard practice in classrooms. Often unable to detect the effect of changes in the short term, and so not informative for rapid improvements.
Improvement (i.e., “practical measures”)	Determining whether an educational change is an improvement.	“If I change my routines for emailing my students, will it create a sense of belonging and promote engagement and learning?"	Very brief and embedded in daily work. Measure only select aspects of constructs that are an intentional focus of improvement work, that are tied to a practical theory, and explicitly signal processes that are actionable by educators. Sensitive to changes.	Primary concern here is predictive validity within classrooms, between classrooms, and/or between schools. Improvements goal is to drive predictive relations with course outcomes to zero.	In many cases practical measures have not yet been developed. Requires building systems (web-based or otherwise) for collecting and rapidly reporting on data. Measures that are relevant in one context may have a different meaning in another.

Table 2

General Uses Cases for Practical Measures

Practical Need	Research Question	Measurement Specification
Assessing changes	Did the change that I implemented actually lead to an improvement?	Repeatable measures that are sensitive to changes over the short term and that reliably predict objective outcomes.
Predictive analytics	Which individuals are highly at risk for the problematic outcome?	Brief, highly predictive measures that are practical to administer and able to be reported on in a timely fashion to enable immediate changes.
Priority setting	Which causes of the problematic outcome continue to be at problematic levels? (And which should be a subsequent improvement priority for the network?)	Brief, highly predictive measures that are practical to administer and able to be reported on before the next improvement cycle.

Table 3

Specific Use Cases for Practical Measures: Examples from Productive Persistence

Practical Need	Example Analysis	Example Action
Assessing changes	Analysis of changes in targeted productive persistence objectives (Figure 2)	Identify interventions that are not showing the changes expected and begin a process of improving them.
Predictive analytics	Develop and deploy an "at-riskness" indicator (Figure 3).	Communicate levels of at-riskness to faculty so that they can quickly deliver targeted supports to students.
Priority setting	Students who are still uncertain about their belonging one month into the course are likely to drop out and fail (even after controlling for other factors in regression analyses) (Figure 5)	Launch sub-networks to design and deliver instructional improvements that address factors identified in priority setting analyses.

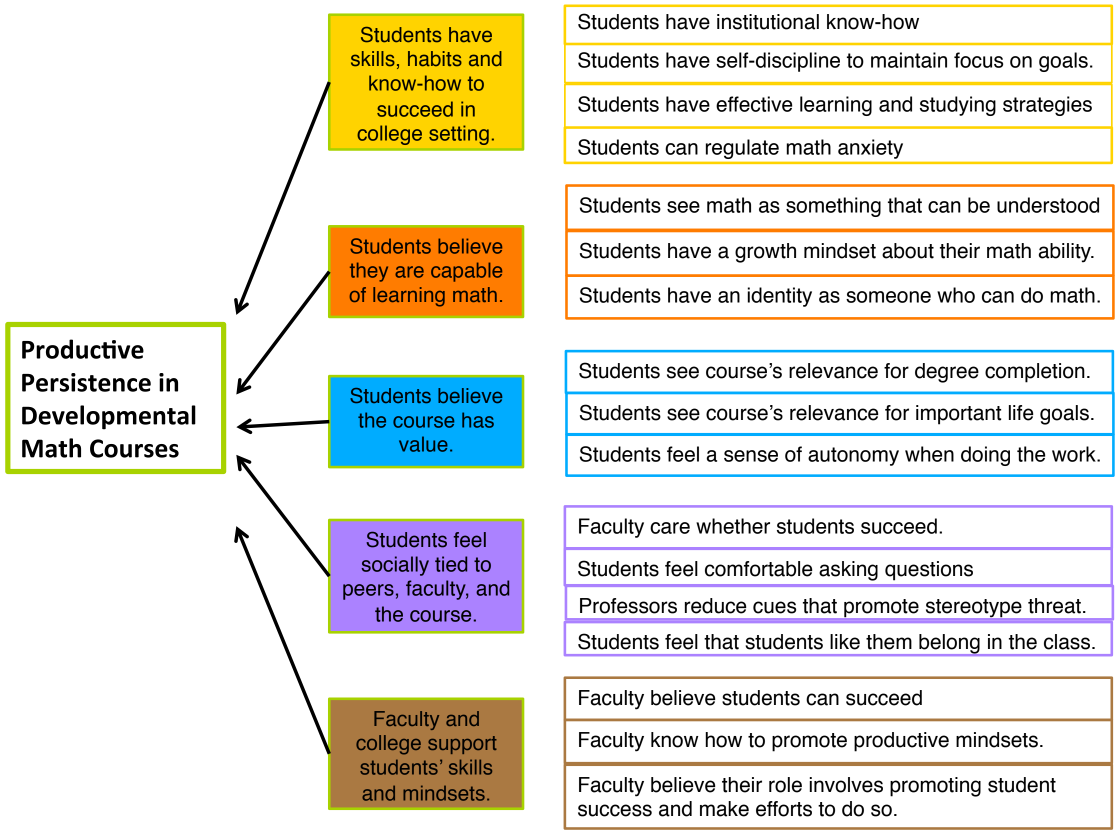

Productive Persistence in Developmental Math Courses
Students have skills, habits and know-how to succeed in college setting.
Students have institutional know-how
Students have self-discipline to maintain focus on goals.
Students have effective learning and studying strategies
Students can regulate math anxiety
Students believe they are capable of learning math.
Students see math as something that can be understood
Students have a growth mindset about their math ability.
Students have an identity as someone who can do math.
Students believe the course has value.
Students see course's relevance for degree completion.
Students see course's relevance for important life goals.
Students feel a sense of autonomy when doing the work.
Students feel socially tied to peers, faculty, and the course.
Faculty care whether students succeed.
Students feel comfortable asking questions
Professors reduce cues that promote stereotype threat.
Students feel that students like them belong in the class.
Faculty and college support students' skills and mindsets.
Faculty believe students can succeed
Faculty know how to promote productive mindsets.
Faculty believe their role involves promoting student success and make efforts to do so.

Figure 1. A framework for "Productive Persistence" (tenacity + effective strategies) in developmental math.

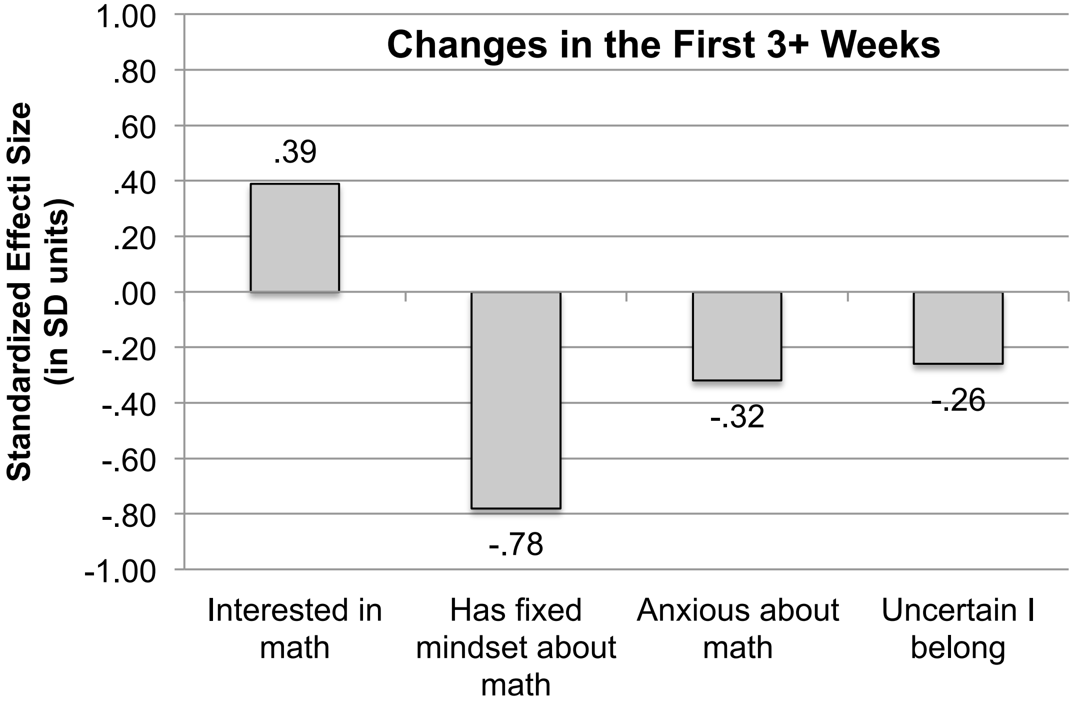
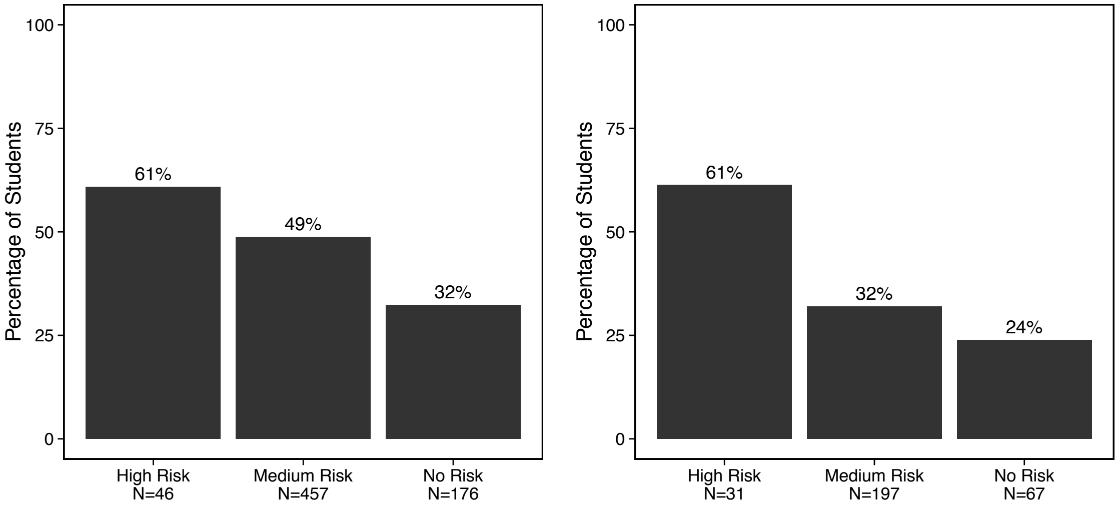


Figure 2. Initial evidence on the efficacy of the productive persistence “Starting Strong” package of activities.

Note. Values show differences between baseline (day 1) and week 3+ values of each cause of productive persistence. All effects significant at $p < .001$.



(a) Statway ($\chi^2(2)\, p < 0.001$) (b) Quantway ($\chi^2(2)\, p = 0.001$)

Figure 3. Productive persistence at-riskness indicator predicts the percentage of students who pass the end-of-term common assessment with a score of 60% or better for Statway and Quantway.

Note. High risk = 3, 4, or 5 productive persistence risk factors; Medium risk = 2 or 1 productive persistence risk factors; No risk = 0 productive persistence risk factors.

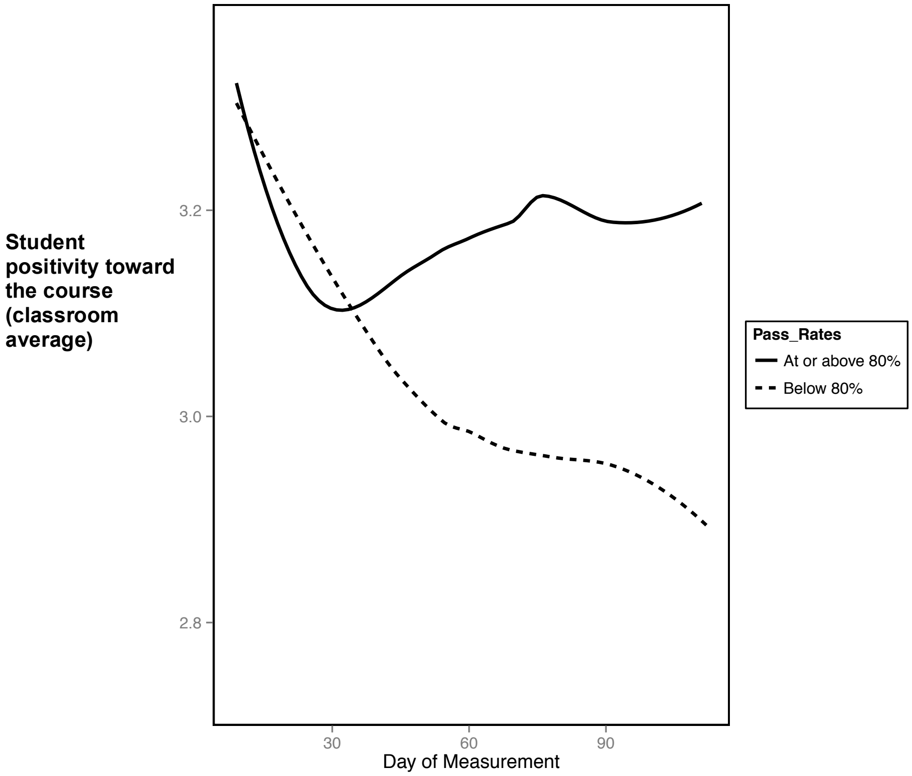


Figure 4. A single-item practical measure of real-time student engagement differentiates classes with high and low course pass rates.

Note. For ratings of positivity, 4 = *Very positive*, 1 = *Very negative*.

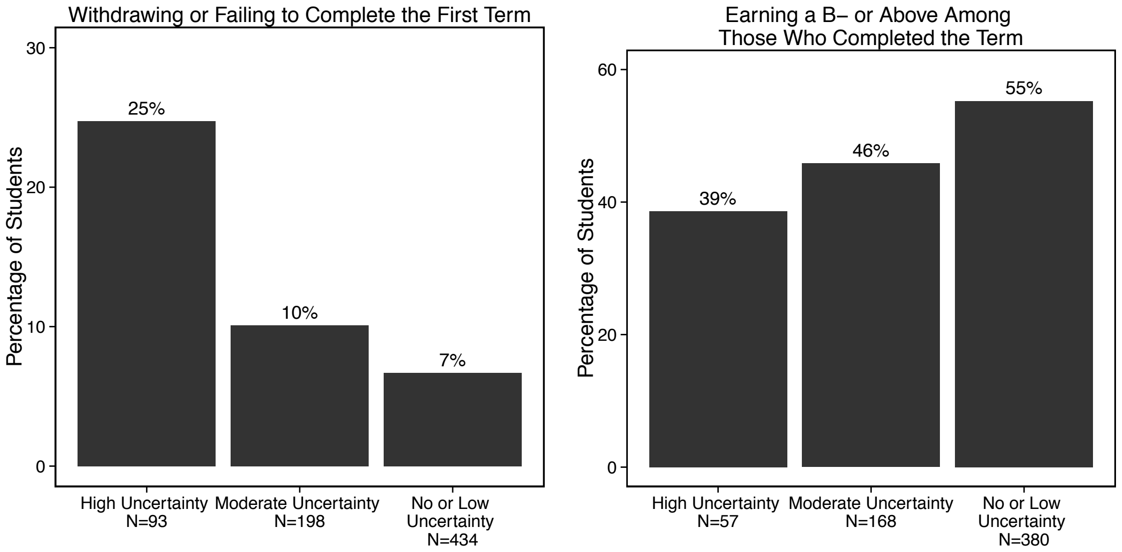


Figure 5. Relation of single-item measure of belonging uncertainty to course outcomes in the Statway® and Quantway®.

Note. Data for Statway® and Quantway® combined. Belonging uncertainty measured at Week 4. Survey item: "When thinking about the Statway® [Quantway®], how often, if ever, do you wonder, 'Maybe I don't belong here?'" Response coding: No or low uncertainty = "*Never*" or "*Hardly ever*"; Moderate uncertainty = "*Sometimes*"; High uncertainty = "*Frequently*" or "*Always*." Withdrawal is indexed by a student's either filing paperwork to officially withdraw from the course or simply failing to attend class and complete work. χ^2 (2) tests, $p < .0001$.

Overview

The "at-riskness" index was developed using data collected from developmental math students in two-year and four-year colleges enrolled in Statway® and Quantway® in the 2011-2012 academic year. Both pathways aim to simplify students' path through their developmental math sequence. Statway integrates developmental mathematical skills and introductory statistics by focusing on data analysis and statistical reasoning. Quantway integrates developmental mathematical skills with quantitative reasoning and literacy.

Four main types of student data were collected:

1. As assessment of mathematical conceptual knowledge prior to the course,
2. "Non-cognitive" skills and mindsets that result in productive persistence,
3. Demographics, and
4. Course performance in Statway or Quantway.

In addition to class meetings, both Pathways make use of an online platform to integrate supplemental content and practice into the course. Other than course performance, the data used were collected through surveys and assessments administered to students via this online platform. Mathematical conceptual knowledge and demographics were collected once on the first day of the course, while indicators of productive persistence were collected twice—once on the first day of the course and again on the fourth week of the course. Students' course performance was collected from institutions upon completion of the term. More detail about the students and the data we collected is described below.

Primary Analytic Sample

The "at-riskness" index was developed using data from the the the first terms Statway and Quantway were taught, the fall of 2011 and the spring of 2012, respectively. A total of 1,077 Statway and 548 Quantway students were enrolled in these terms. For this analysis we used data from 950 Statway students enrolled in the fall of 2011 and 441 Quantway students. These were the students for whom enough of the key data had been collected to perform the necessary analyses. See the missing data section for more details about how students were excluded from the analytic sample.

The vast majority of Statway students (78 percent) placed at least two levels below a college-level mathematics course and almost half would be required to take at least one developmental reading course as well. About 60 percent of the students are female and less than 30 percent were raised in families where the mother held a college degree. Over two-thirds of the Statway students are minorities and 45 percent grew up in an environment where a language other than English was spoken.

Similarly, over half (56 percent) of Quantway students placed into mathematics courses two levels or more below college-level mathematics and 39 percent placed into developmental reading course as well. Sixty percent of Quantway students are female and about one third came from families where the mother obtained a college degree. A small percentage of students (13 percent) grew up in a home where a language other than English was spoken.

Overall, both Pathways enroll traditionally underserved populations but differ slightly in their ethnic compositions, likely reflecting the differences in the states in which the two Pathways are located. In Statway, 33 percent of students are Hispanic, 30 percent are Caucasian, and 25 percent are African. A small percentage are Asian and other ethnic minorities. Quantway students are predominantly Caucasian (42 percent) and African American (41 percent) with smaller percentages of other ethnic minorities.

In Statway, approximately 93% of students of complete the first term, with approximately 70% of all students passing. In Quantway, approximately 82% of students of complete the first term,and about 56% of all students passed that term.

Replication Sample

For replication analyses, we used data from students enrolled in Statway and Quantway in the fall of 2012. This group consists of 777 Statway and 440 Quantway students who, by and large, come from the same colleges, have the same instructors, and have many of the same characteristics as the 2011 students.

Measures

Mathematical Conceptual Knowledge

Upon creating an account for the online component of Statway and Quantway, students are immediately directed to an online survey about their background. One component of this background survey is a baseline assessment of their mathematical knowledge prior to beginning the course. It is designed to be an assessment of students' conceptual understanding of basic mathematical ideas rather than of their computational skills. This 42 item test takes students between 10 and 60 minutes to complete, has a median completion time of 19 minutes, and a median score of 24 correct responses.

For the purpose of the "at-riskness" index, students' scores are classified as "at risk" and "not at risk" in terms of baseline math conceptual knowledge. These categories are defined by scoring less than a 22 and scoring a 22 or greater, respectively. The relationship between scores on this assessment, continuous and dichotomous, and course outcomes can be visualized in Figure 1.

Sample items from the assessment of math conceptual knowledge

Is 0.3% equivalent to 0.03?
○ Yes
○ No

Is 2^3 equivalent to 6?
○ Yes
○ No

A pound of apples costs $1.98 per pound. To find the cost of 0.75 pounds of apples, which of these calculations would you use?
○ 1.98×0.75
○ $1.98 \div 0.75$
○ $1.00 - 0.75$

○ $1.98 - 0.75$

$$n + \frac{1}{3} = a$$

In the equation above, n and a are positive numbers. Which of the following is true about a and n?

○ a is greater than n
○ a is less than n
○ a is equal to n
○ It's impossible to tell

Which of the following numbers is between $\frac{5}{6}$ and 1?

○ $\frac{6}{5}$
○ $\frac{5}{7}$
○ 0.55
○ 0.9
○ There is no number between $\frac{5}{6}$ and 1

Is $n \div \frac{1}{2}$ equivalent to half of n?

○ Yes
○ No

Productive Persistence

Another component of the background survey students respond to on their first login to the platform is a series of items designed to tap into students' skills and mindsets that result in productive persistence. The survey items used to construct the "at-riskness" index are listed below, grouped into five indicators. The "at-riskness" index is the number of these indicators for which a student expresses problematic responses. Being "at-risk" for each of these five risk factors is determined by empirically-derived cutpoints based on the relationship between the indicator and performance on a summative assessment of course content. The relationship between the productive persistence indicators, continuous and dichotomous, and course outcomes can be visualized in Figures 2 through 6.

Note that students are directed to the productive persistence survey again approximately four weeks into the course. This allows for the analysis of changes in students' levels on the productive persistence drivers. The change data can be used, for example, to assess the quality of implementation of interventions designed to affect these drivers.

Items used to create math and statistics anxiety measure

The following four items were used to create a measure of math and statistics anxiety. This composite was calculated by taking the unweighted average of a student's responses to these questions. If a student did not answer all four of these items, their anxiety composite score is the average of the items they did respond to.

The items below refer to things that may cause fear or tension. There are no right or wrong responses, only the way you feel about the statement. Don't think about it too much: just mark the response that first comes to mind.

How anxious would you feel listening to a lecture in math or statistics class?
○ Extremely anxious (1)
○ Very anxious (2)
○ Moderately anxious (3)
○ Slightly anxious (4)
○ Not at all anxious (5)

How anxious would you feel taking a math or statistics test?
○ Extremely anxious (1)
○ Very anxious (2)
○ Moderately anxious (3)
○ Slightly anxious (4)
○ Not at all anxious (5)

How anxious would you feel signing up for a course in statistics and probability?
○ Extremely anxious (1)
○ Very anxious (2)
○ Moderately anxious (3)
○ Slightly anxious (4)
○ Not at all anxious (5)

How anxious would you feel the moment before you got a math or statistics test back?
○ Extremely anxious (1)
○ Very anxious (2)
○ Moderately anxious (3)
○ Slightly anxious (4)
○ Not at all anxious (5)

Items used to create "fixed mindset" measure

The following four items were used to create a measure of having a "fixed mindset." A composite of these items was calculated by transforming each response onto a continuous 0 to 1 scale, taking the unweighted average of the transformed responses, and then transforming the averages back to a continuous scale ranging from 1 to 6. If a student did not answer all four of these items, their "fixed mindset" composite is the average of their responses to the items they did answer.

Read the statement below and mark how much you agree or disagree with it: 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.
○ Strongly agree (1)
○ Agree (2)
○ Mostly agree (3)

○ Mostly disagree (4)
○ Disagree (5)
○ Strongly disagree (6)

The next few questions will ask you to think about what normally goes through your mind when you get high or low scores in a math or statistics class.

First, imagine that you got a high grade on a math test or quiz. How true or not true would you normally think each of these statements was?

I got lucky.
○ Extremely true (1)
○ Very true (2)
○ Moderately true (3)
○ Slightly true (4)
○ Not at all true (5)

I studied the right way.
○ Extremely true (1)
○ Very true (2)
○ Moderately true (3)
○ Slightly true (4)
○ Not at all true (5)

Now, imagine that you got a low grade on a math test or quiz. How true or not true would you normally think this statement was?

I'm not smart enough at math.
○ Extremely true (1)
○ Very true (2)
○ Moderately true (3)
○ Slightly true (4)
○ Not at all true (5)

Item used to measure "belonging uncertainty"

When you think about your college, how often, if ever, do you wonder: "Maybe I don't belong here?"
○ Always (1)
○ Frequently (2)
○ Sometimes (3)
○ Hardly ever (4)
○ Never (5)

Item used to measure “stereotype threat”

Do you think other people at your school would be surprised or not surprised if you or people like you succeed in school?

○ Extremely surprised (1)
○ Very surprised (2)
○ Slightly surprised (3)
○ Moderately surprised (4)
○ Not surprised at all (5)

Measure of “grit”

This unobtrusive behavioral measure of “grit” is a dichotomous variable, which indicates whether or not a student completed the last problem on the assessment of math conceptual knowledge, given that he or she completed the first problem.

Demographics

The final component of the background survey that students are administered on the first day is series of demographic questions. From these student self-reports, we collect the following information:

- The student’s income,
- Whether or not the student is African American,
- Whether or not the student is Hispanic / Latino,
- Whether or not the student grew up in a home where English was the only language spoken,
- Whether or not the student has dependents, and
- Whether or not the student’s mother earned a college degree.

Course Performance

Two measures of course performance were collected to assess student success in Quantway and Statway: course grade and performance on a common assessment administered at the end of the term. Official course grades are collected at the completion of the term from the institutions that students attend. Based on grades earned in the first term of Quantway and Statway, a variety of student success variables were examined. First, student grades can be translated to a four point GPA scale with an “F” corresponding to zero points and an “A” corresponding to four points. Students who did not complete the course, i.e. withdrew, dropped, or received a grade of incomplete, are excluded in analyses which examine grade on a four point scale. We also created a dichotomous course grade variable indicating whether or not a student earned a B- or above. This grade was selected as the minimum grade a student would need to earn in order to have mastered enough of the material to be successful in future college level math courses. Note that for analyses which rely on course grade data of earning a B- or above, not all of the Quantway data can be used. This is because for some of the schools in Quantway, “P” is the only passing grade that is assigned. In these schools, Quantway is graded on a Pass/No Pass basis because it is a developmental course. Finally, we created a course withdrawal variable using students’ grades. In our analyses, course withdrawal consists of students who officially withdrew from the course, i.e. earned a grade of “W,” as well as students who unofficially withdrew. Stunts who unofficially withdrew are those students who received grades of incomplete, or other institution-specific grades indicating that they stopped attending class and the instructor withdrew them from the course.

Data about assessment performance is collected from students' Quantway and Statway instructors. The instructors administer the assessment and have students fill in answer sheets. The student answer sheets are then returned to and scored at the Carnegie Foundation. In Quantway, the common test was a 40 item assessment of students' quantitative reasoning abilities using methods typically covered in a beginning algebra course. The mean score for this test was a 66%. In Statway, the common assessment was a 44 item test designed to measure students' data analysis and causal reasoning skills that are taught in a typical introductory statistics course. The mean score for this assessment was a 62%. Note that in the 2011-2012 academic year from which the data was collected, Quantway was only a one term course as the second term was still being developed. Therefore, for comparison purposes, nearly all analyses were conducted with first term course performance measures from Statway and Quantway. In most analyses, students' assessment scores were classified as failing and passing, with failure being defined as below 60% correct. This threshold was chosen because students who earned a C+ as a course grade earned approximately a 60%, on average, on the common assessment.

For some analyses though, such as the regression model described in Table 6, assessment performance was treated as a continuous outcome variable measured in logits (log odds units). Students' correct and incorrect responses to the assessment items are transformed into scores in logits through the Rasch model. The Rasch model produces an interval scale that determines ability measures for students, called *person measures*, and difficulty measures for assessment items, called *item difficulties*. Crucially, the Rasch model puts person measures and item difficulties on a common scale. Items are arranged on the scale according to how likely they are to be answered correctly. Students are then placed on the scale based on which items they answered correctly. Person measures and item difficulties are assigned based on where students and items fall on the scale. More difficult items and students who answer more questions correctly appear at the top of the scale while easier items and students answering fewer items correctly appear at the bottom. The person measures are chosen over raw scores or percent correct as the assessment performance outcome variable. This is because the measures form an interval scale through which students' abilities can be directly and meaningfully compared to specific assessment items and can be calculated regardless of which items students take. Note though, that in Quantway and Statway the Rasch measures are highly correlated with percent correct ($r(335) = .97, p < .001$ and $r(720) = .99, p < .001$, respectively).

Finally, one course success measure was created that captures both course grade and common assessment performance. Students who passed the common assessment and earned a B- or above were considered successful on this variable and students who either failed the assessment, earned below a B-, or both were considered unsuccessful.

Missing Data

As previously described, our analyses rely on data about students' math conceptual knowledge, levels of productive persistence measures, and demographics, which are collected through surveys and assessments on the online component of Statway and Quantway. Missing data for these three components occurs when students choose to skip survey or test items. In general, students tended to have close to complete data, or nearly no data on the "at-riskness" index variables of interest. We chose to eliminate student records where data on seven or more of the nine "at-riskness" index variables is missing. Afterwards, 441 valid Quantway student records and 950 valid Statway student records remain. Table 1 shows what percentage of the remaining student records are missing zero through six of the nine variables of interest.

Table 1: Percentage of Student Records Missing Data on Zero Through Six "Risk Factors" in Quantway and Statway.

	0	1	2	3	4	5	6
Quantway	90%	7%	1%	< 1%	< 1%	< 1%	1%
Statway	87%	9%	1%	< 1%	< 1%	< 1%	2%

Further description of the patterns of missing data are described below.

Missing Math Background Scores

Math background scores are only missing when students do not answer *any* of the 42 items on the baseline assessment. Approximately 3% of Quantway students and 4% of Statway students have missing math conceptual knowledge scores because they skipped the assessment entirely.

Missing Productive Persistence Data

The productive persistence "at-riskness" index is based on students' responses to five productive persistence indicators, composed of a total of 10 survey items and 1 behavioral measure of "grit." To be missing on any of the five productive persistence indicators would mean that a student did not respond to the survey item(s) which make up the indicator. Recall that the mathematics anxiety measure and the "fixed mindset" measure are averages of four items each. When students did not respond to all four items, their composite score is the average of their responses to the items they did answer. This means that only students who did not respond to any of the four questions have missing for mathematics anxiety or "fixed mindset." Students are missing data on the other three "risk factors" only if they are missing responses to the single item that each "risk factor."

Table 2 shows the percentage of students who are missing data for each of the five productive persistence "risk factors." Table 3 displays the percentage of students in each course who are missing data on 0 through 5 "risk-factors." All students' "non-cognitive at-riskness" totals are the count of the number of these 5 factors for which they are "at-risk." For students missing data on at least one of the factors, the maximum their "non-cognitive at-riskness" totals can be is the number of "risk-factors" for which they have data.

Table 2: Percentage of Data Missing for each Productive Persistence "Risk Factors" in Quantway and Statway.

	Math Anxiety	Fixed Mindset	Belonging Uncertainty	Stereotype Threat	Grit
Quantway	< 1%	< 1%	1%	2%	0%
Statway	< 1%	< 1%	3%	3%	0%

Missing Demographics Data

Table 4 shows the percentage of students missing each piece of demographic data. Table 5 displays the percentage of students in each Pathways course who are missing data on 0 through 3

Table 3: Percentage of Students Missing Data on 0 through 4 Productive Persistence "Risk Factors" in Quantway and Statway.

	0	1	2	3	4	5
Quantway	98%	1%	1%	0%	0%	0%
Statway	97%	1%	2%	0%	0%	0%

demographic "risk-factors."

Table 4: Percentage of Data Missing for Each Piece of Demographic Data in Quantway and Statway.

	Income	African American	Hispanic/Latino
Quantway	32%	5%	5%
Statway	33%	10%	10%

	Home Language	Dependents	Maternal Ed.
Quantway	2%	2%	5%
Statway	3%	3%	7%

Table 5: Percentage of Students Missing Data on 0 through 3 Demographic "Risk Factors" in Quantway and Statway.

	0	1	2	3
Quantway	91%	7%	1%	1%
Statway	87%	10%	0%	3%

Missing Course Outcomes

As outcome variables, we use data about common assessment performance and course grade. Common assessment performance data is missing if students did not take the common assessment or if their response forms were not returned to the Carnegie Foundation. In both Pathways, some instructors did not return student answer keys for their entire class. This means that we cannot determine assessment performance for the students in those classes. Excluding those courses, in Quantway and Statway respectively, assessment scores are missing, however. These additional missing assessment scores come from students who did not take the assessment even through response sheets were returned for other students in their class. These are the students that failed to show up for the test. In Quantway, there are 146 missing assessments, 130 of which are the students who failed to show up for the test. In Statway, there are 271 missing assessments, 167 of which are the students who failed to show up for the test.

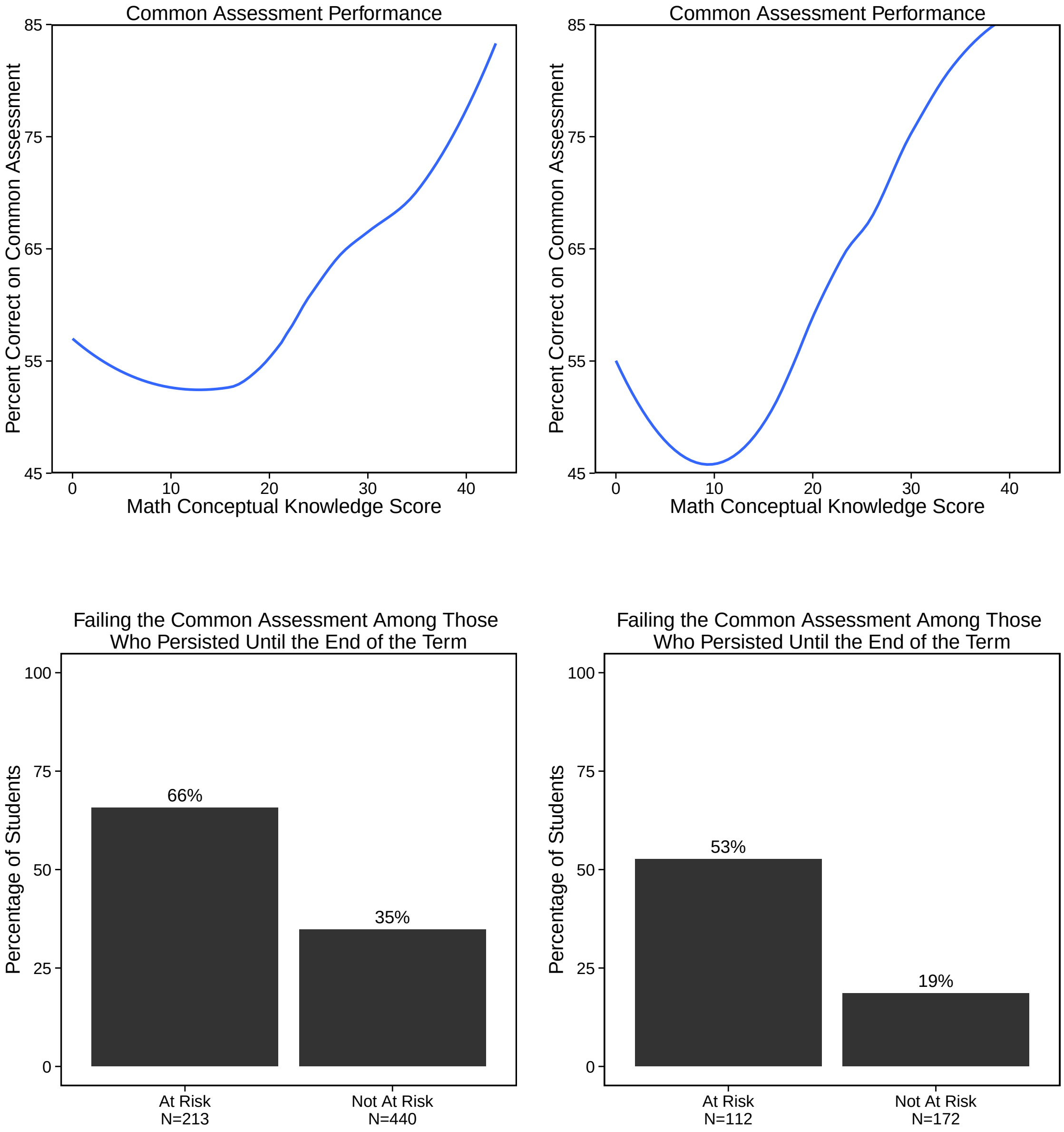
Regardless of the reason their data is missing, when students do not have assessment scores, they are excluded from analyses where the outcome variable is performance on the common assessment.

Students' course grades are missing when institutions do not report them or when the student drops the course before the first census date. Students missing grade data are excluded from analyses where the outcome variable is course grade. In Statway, there are 286 student records missing grades, 36 of which are the students who left before the first census and therefore did not receive a grade, even though grades were reported for other students in their institution. In Quantway, there are no missing grades do to institutions failing to report for all of their Quantway students. Instead, there are only 27 missing grades from students who dropped before the first census.

Analyses

Choosing Cutpoints to Create Dichotomous "Risk Factor" Variables

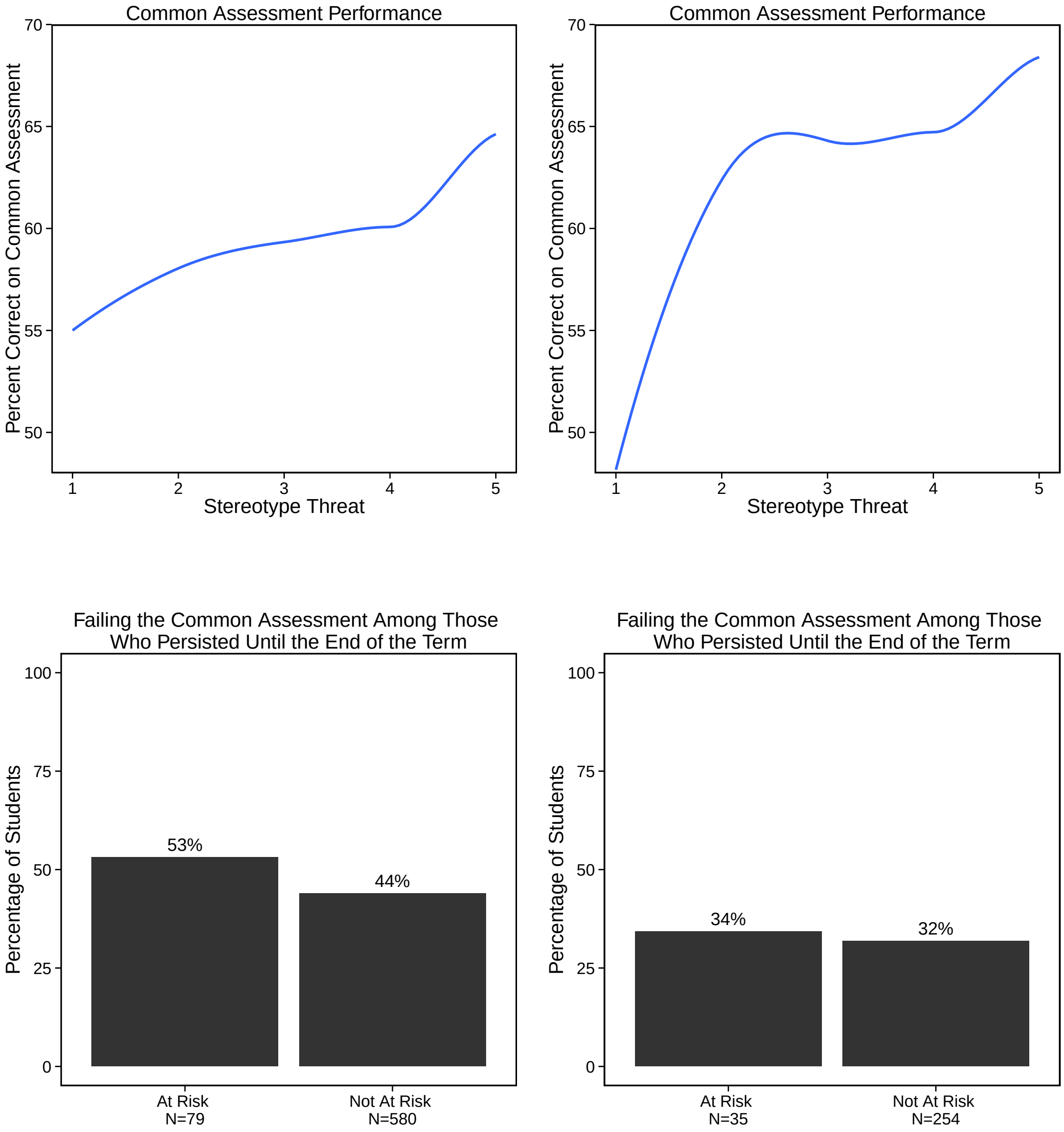
The following figures show the relationship between student background variables and performance on the common assessment in Statway and Quantway. For the continuous student background variables, the relationship between the continuous background variable and percent correct is displayed in the top panel. From this graph a cut-point is derived and is used to classify student responses on that variable as "at risk" or "not at risk." The bottom panel shows the percentage of students in each risk category who fail the common assessment. For naturally dichotomous student background variables, only the figures depicting the percentage of students in each group who fail the common assessment can be created. Students are only included in these graphs if they have data on the background variable and an assessment score.



(a) Statway ($\chi^2(1)\, p < 0.001$)

(b) Quantway ($\chi^2(1)\, p < 0.001$)

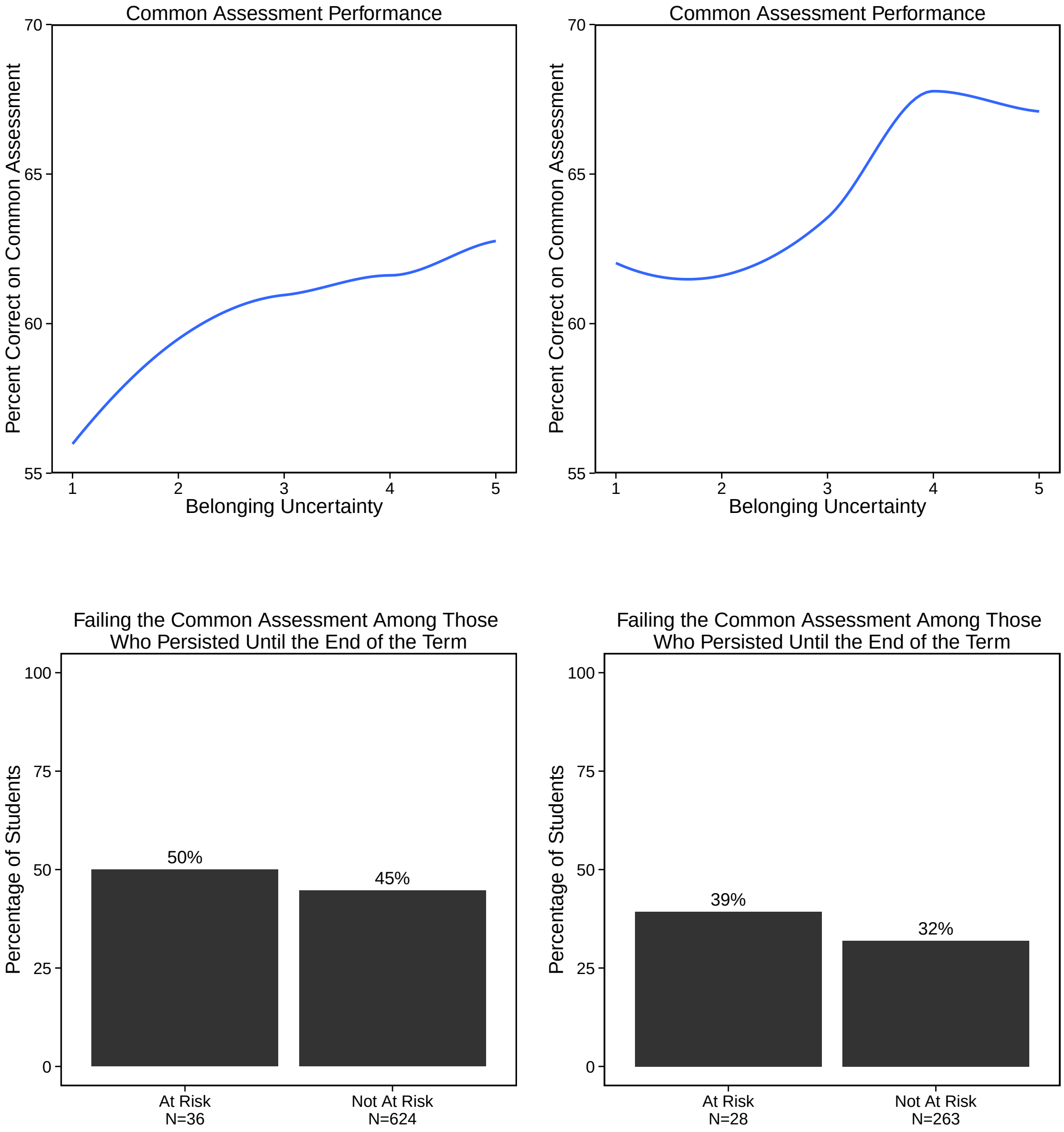
Figure 1: Math conceptual knowledge and failing the common assessment. At risk: score < 22; not at risk: score ≥ 22.



(a) Statway ($\chi^2(1)\, p = 0.123$)

(b) Quantway ($\chi^2(1)\, p = 0.776$)

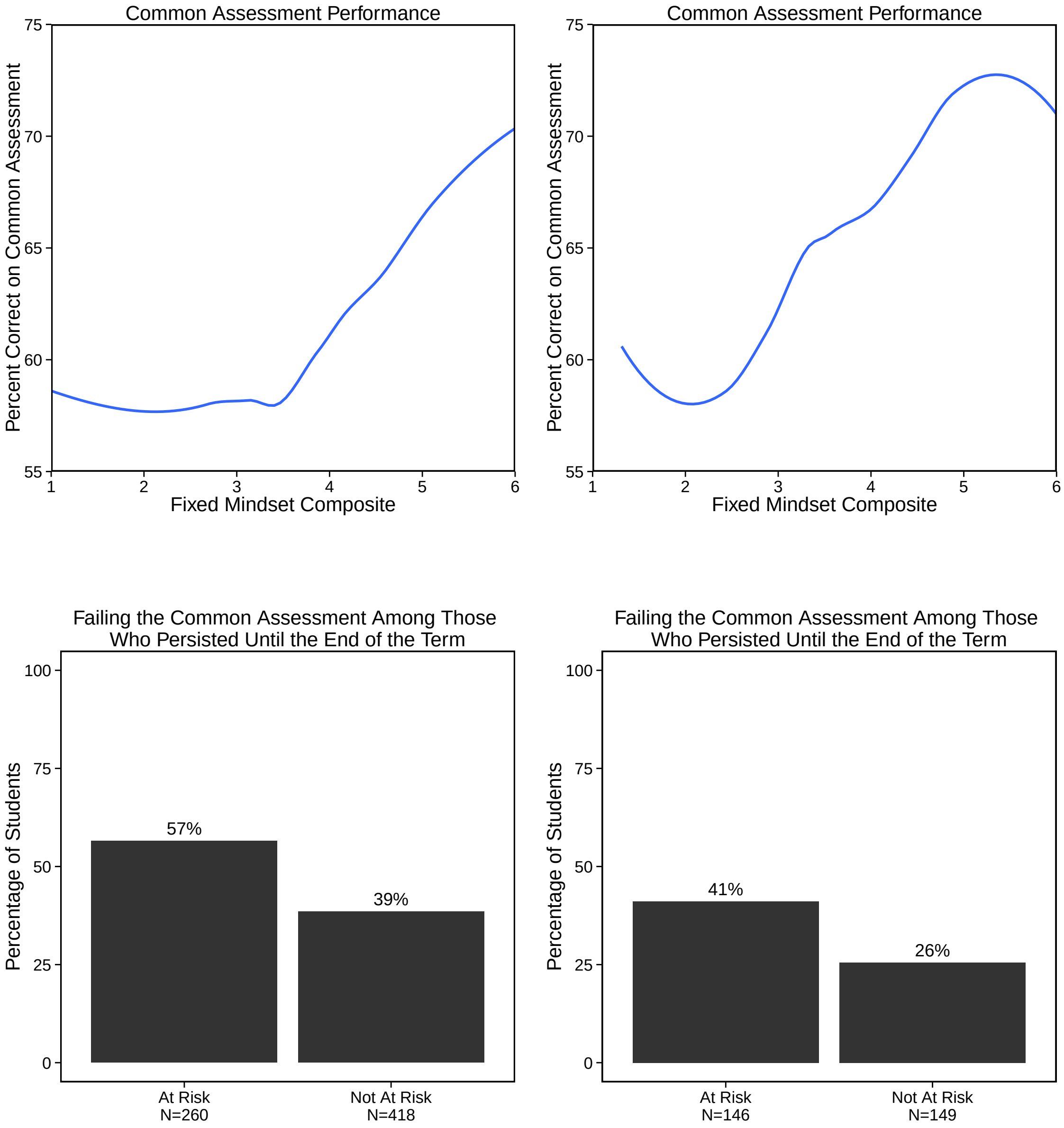
Figure 2: Stereotype threat and failing the common assessment. At risk: response = 1 or 2; not at risk: response = $3, 4$ or 5.



(a) Statway ($\chi^2(1)\, p = 0.535$)

(b) Quantway ($\chi^2(1)\, p = 0.431$)

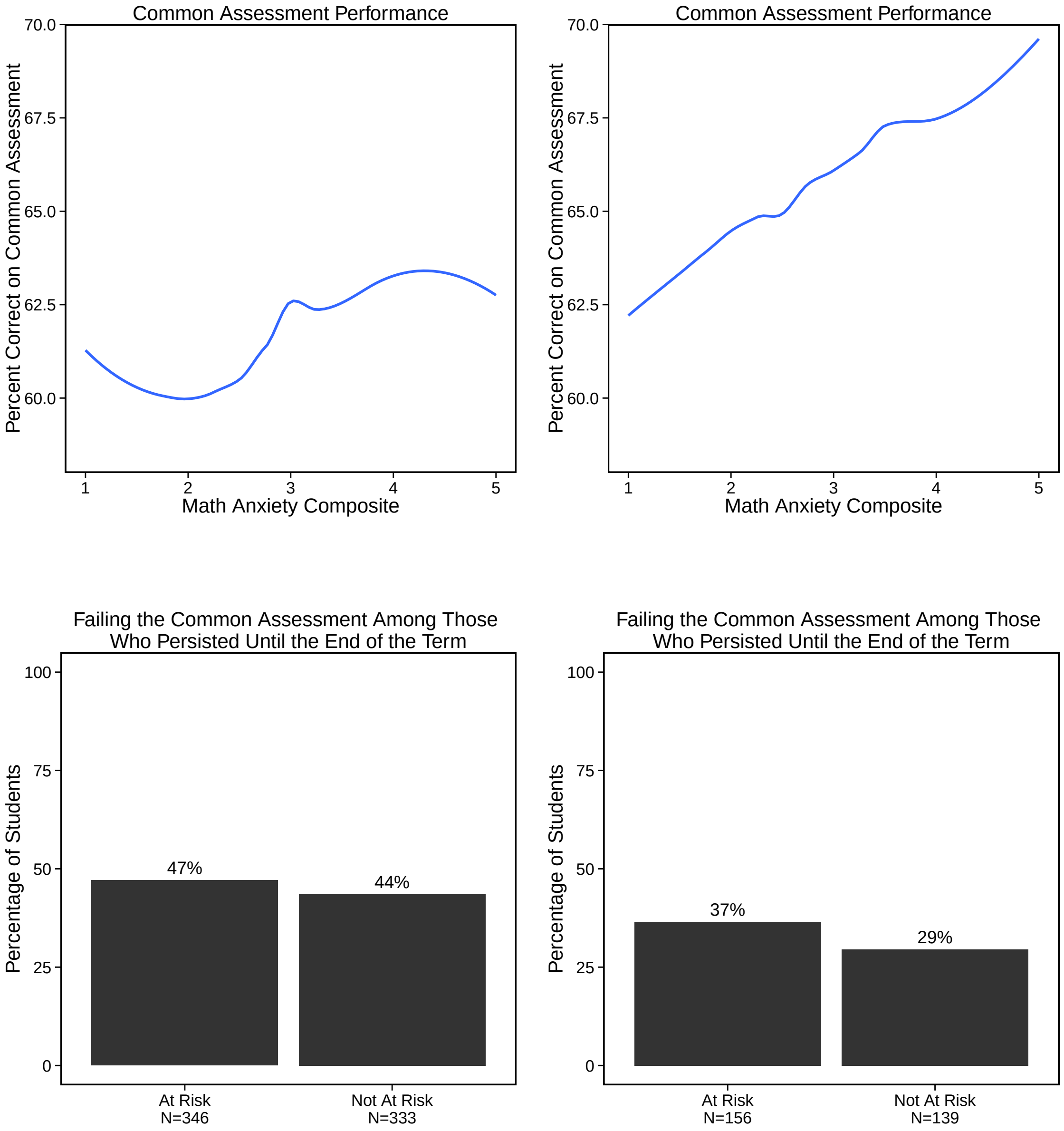
Figure 3: Belonging uncertainty and failing the common assessment. At risk: response = 1 or 2; not at risk: response $= 3, 4$ or 5.



(a) Statway ($\chi^2(1)\, p < 0.001$)

(b) Quantway ($\chi^2(1)\, p = 0.004$)

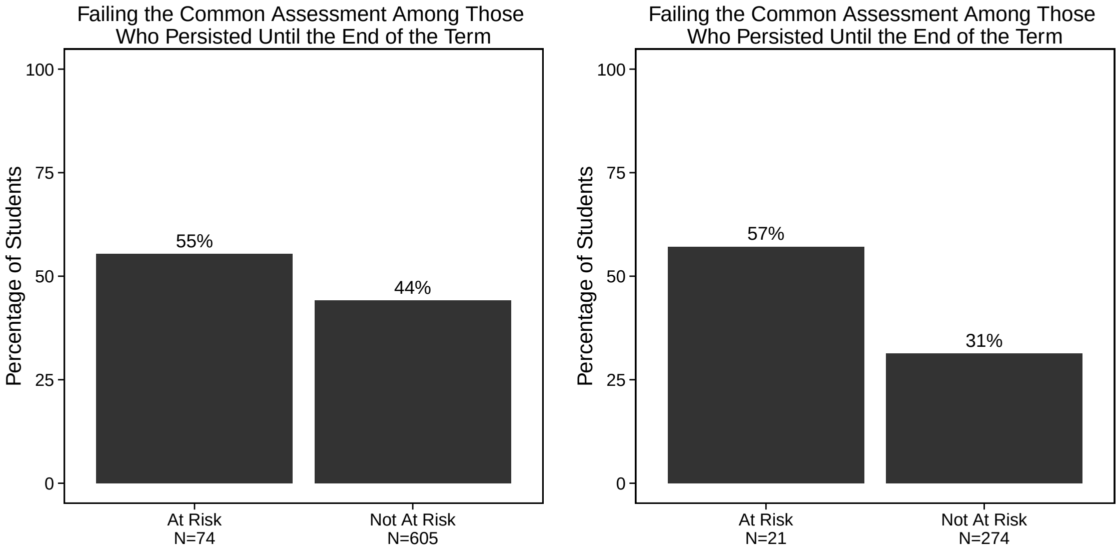
Figure 4: Fixed mindset and failing the common assessment. At risk: score ≤ 3.5; not at risk: score > 3.5.



(a) Statway ($\chi^2(1)\, p = 0.351$)

(b) Quantway ($\chi^2(1)\, p = 0.2$)

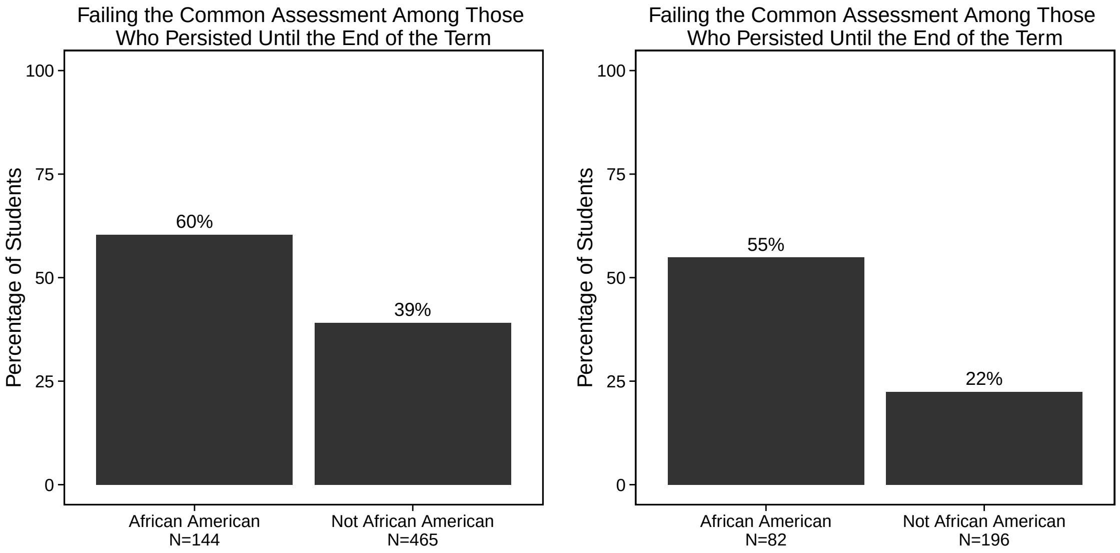
Figure 5: Math/statistics anxiety and failing the common assessment. At risk: response = 1 or 2; not at risk: response $= 3, 4$ or 5.



(a) Statway ($\chi^2(1)\, p = 0.066$)

(b) Quantway ($\chi^2(1)\, p = 0.016$)

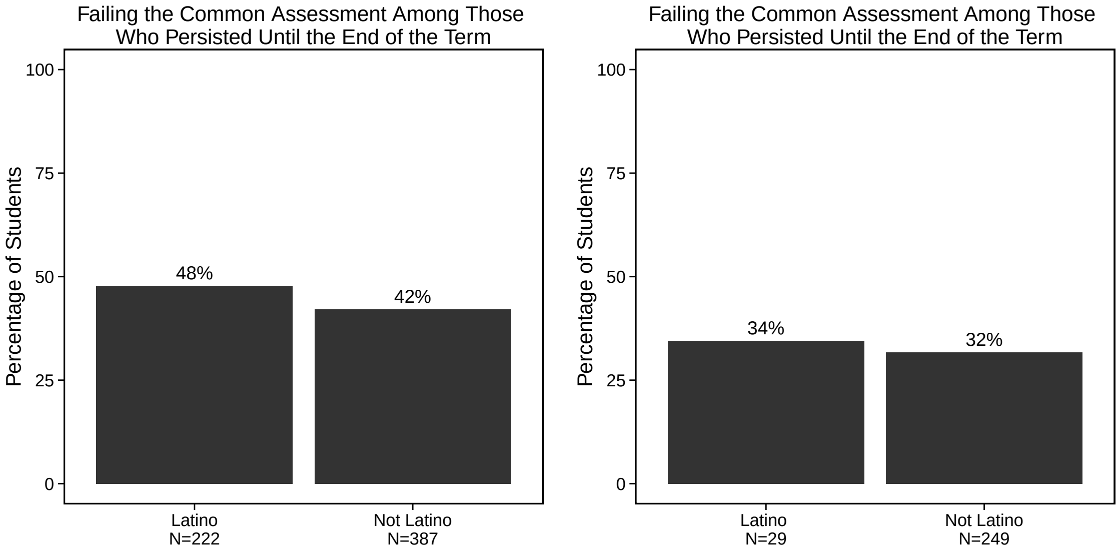
Figure 6: "Gritty" behavior and failing the common assessment. At risk: assessment not completed; not at risk: assessment completed.



(a) Statway ($\chi^2(1)\, p < 0.001$)

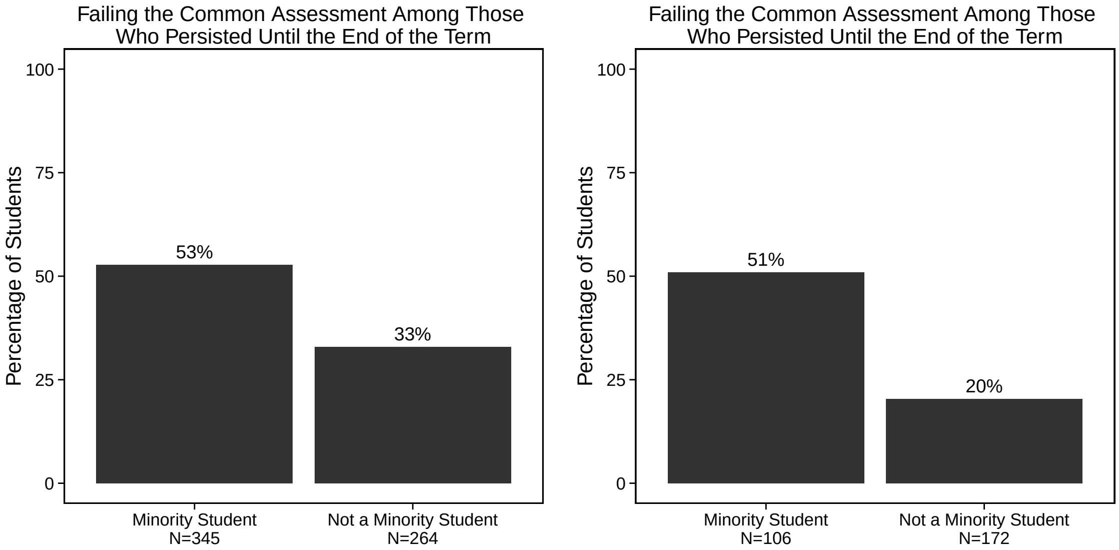
(b) Quantway ($\chi^2(1)\, p < 0.001$)

Figure 7: African American student achievement gap on the common assessment.



(a) Statway ($\chi^2(1)\, p = 0.178$) (b) Quantway ($\chi^2(1)\, p = 0.763$)

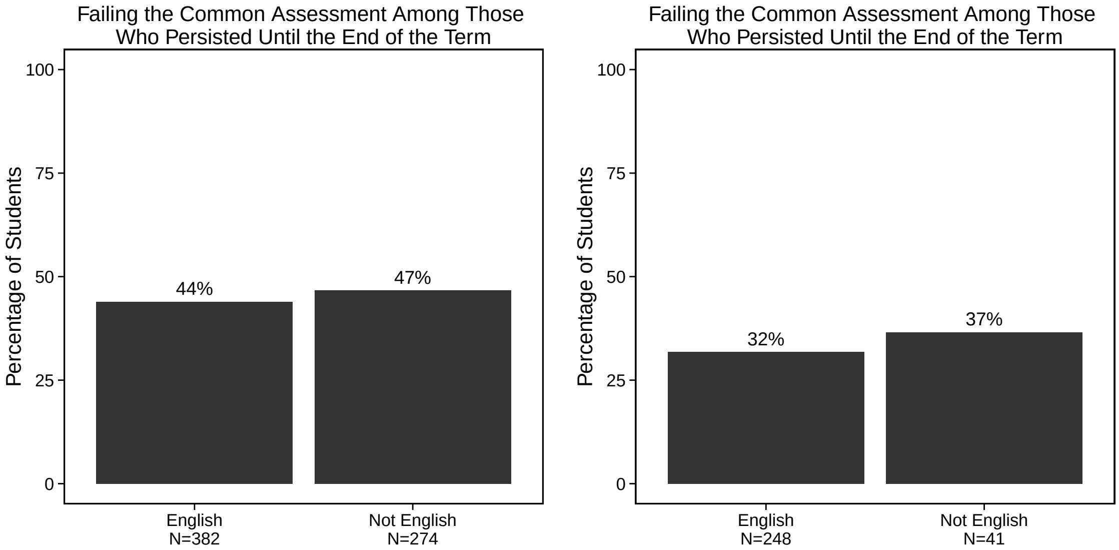
Figure 8: Latino student achievement gap on the common assessment.



(a) Statway ($\chi^2(1)\, p < 0.001$)

(b) Quantway ($\chi^2(1)\, p < 0.001$)

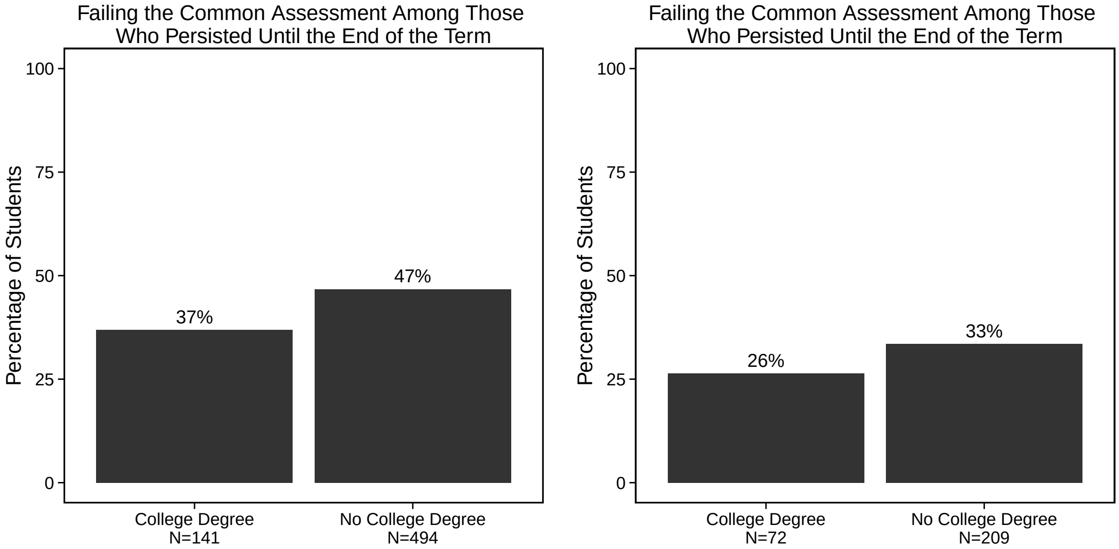
Figure 9: Minority achievement gap on the common assessment. Minority category consists of African American and Latino students.



(a) Statway ($\chi^2(1)\, p = 0.487$)

(b) Quantway ($\chi^2(1)\, p = 0.549$)

Figure 10: Language spoken at home and failing the common assessment.



(a) Statway ($\chi^2(1)\, p = 0.037$)

(b) Quantway ($\chi^2(1)\, p = 0.264$)

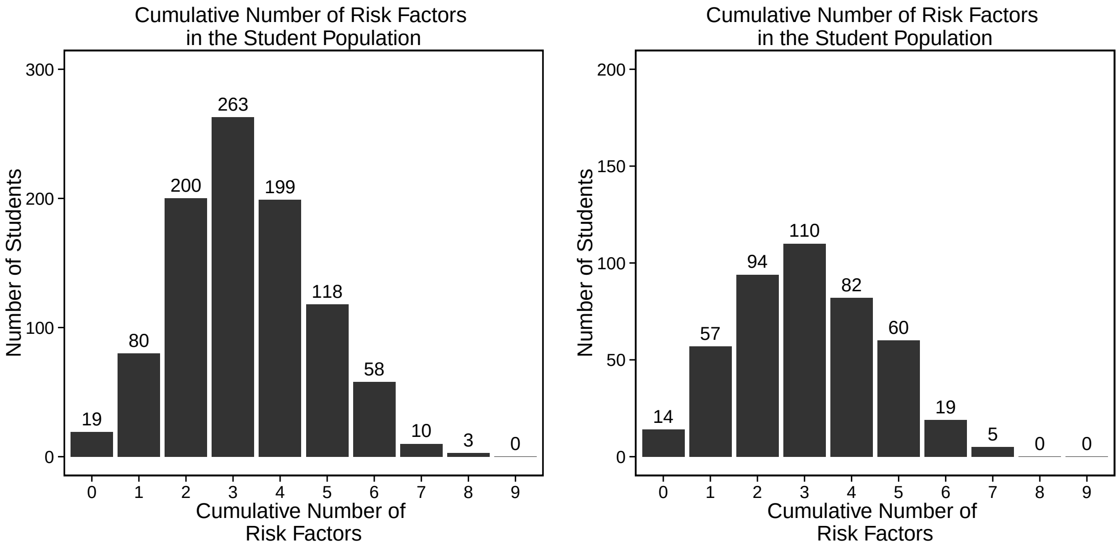
Figure 11: Maternal education and failing the common assessment.

Cumulative Math Background, Productive Persistence, and Demographic Risk

From all of the student background variables collected, nine dichotomous "risk factors" were created (one math conceptual knowledge, five productive persistence, and three demographic factors):

1. Math conceptual knowledge,
2. Math/statistics anxiety,
3. "Fixed mindset,"
4. "Belonging uncertainty,"
5. "Stereotype threat,"
6. "Grit,"
7. Minority status (being African American or Hispanic / Latino),
8. Language spoken at home, and
9. Maternal education.

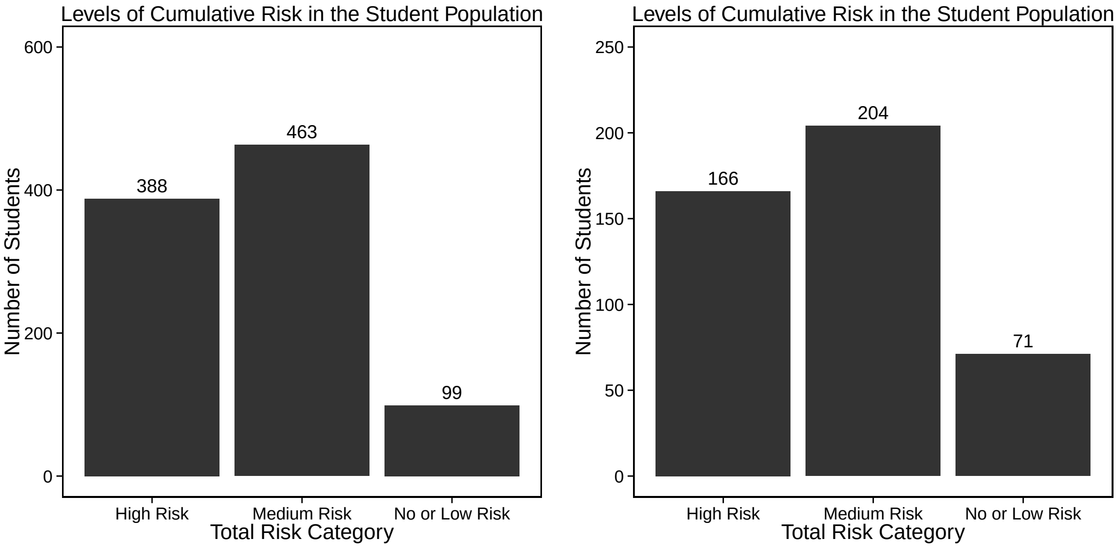
The number of these factors for which a student is classified as "at risk" are counted to determine a cumulative "at-riskness" measure, ranging from 0 to 9. The "at-riskness" totals can be reasonably grouped into three categories: low, medium, and high total risk. The following figures display the number of students in these risk categories and how the categories relate to different course outcomes.



(a) Statway

(b) Quantway

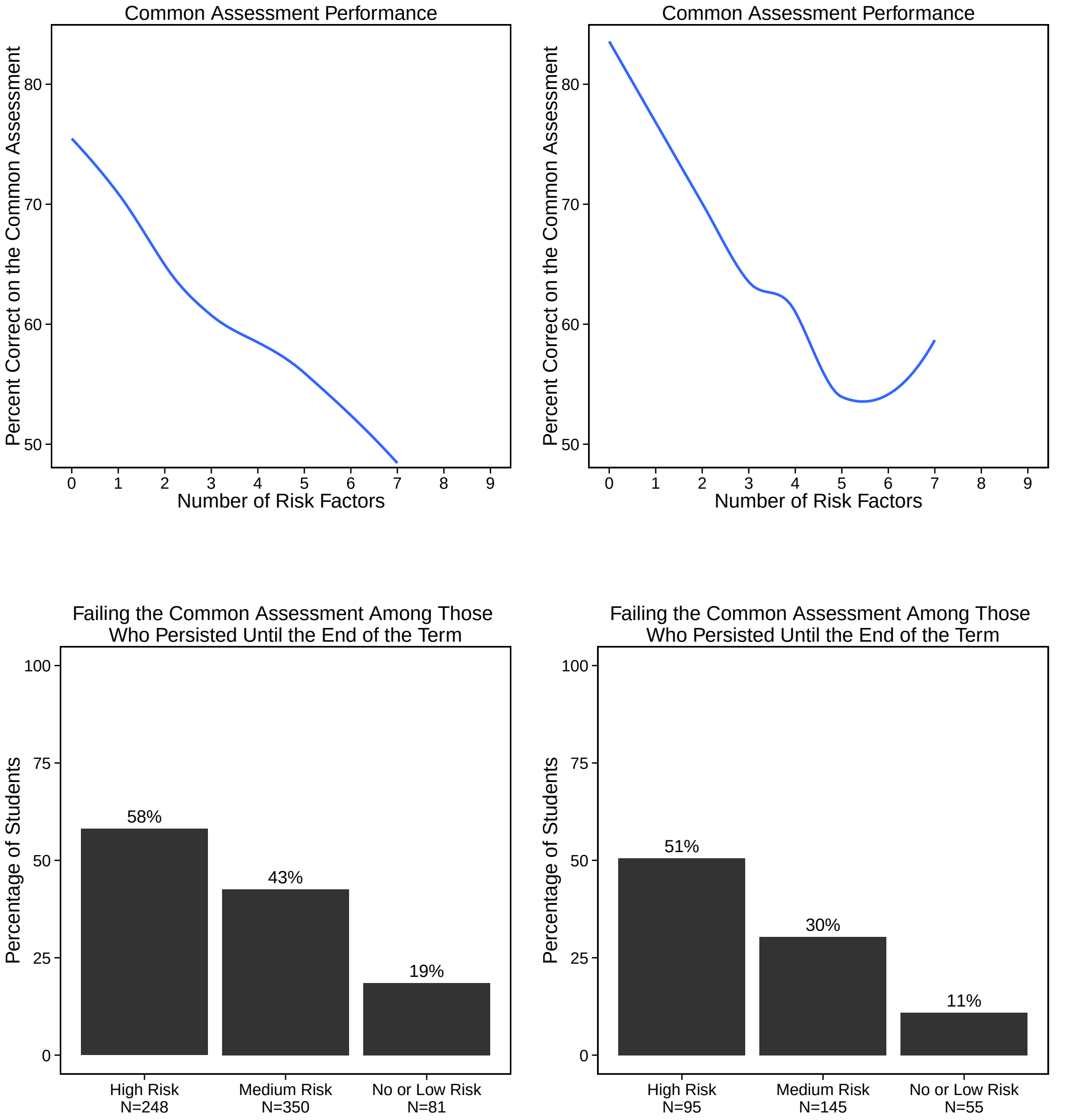
Figure 12: Cumulative number of math, non-cognitive, and demographic risk factors.



(a) Statway

(b) Quantway

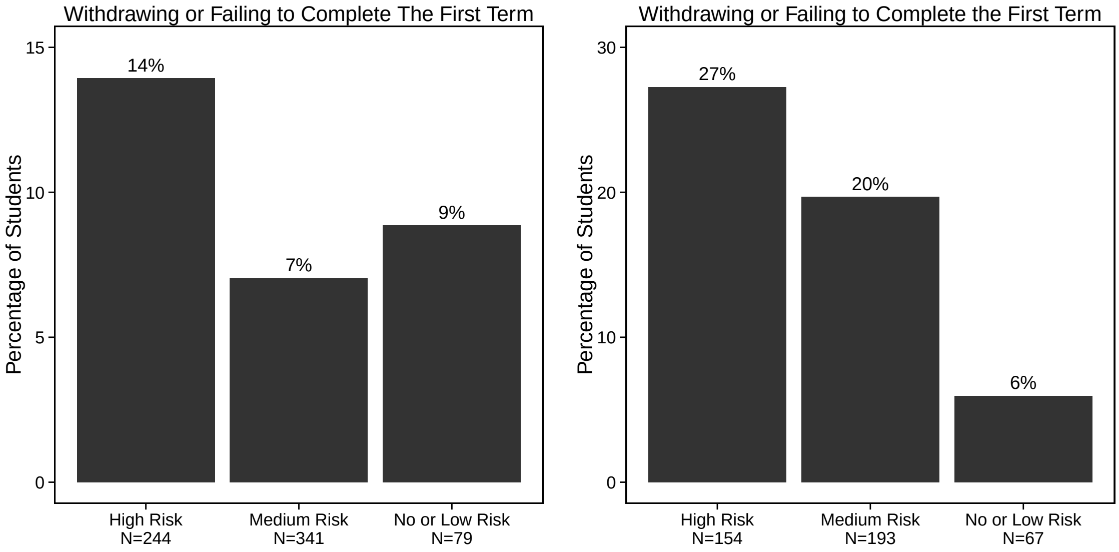
Figure 13: Cumulative Risk Category. High risk: 4 through 9 risk factors; medium risk: 2 or 3 risk factors; no or low risk: 0 or 1 risk factors.



(a) Statway ($\chi^2(2)\, p < 0.001$)

(b) Quantway ($\chi^2(2)\, p < 0.001$)

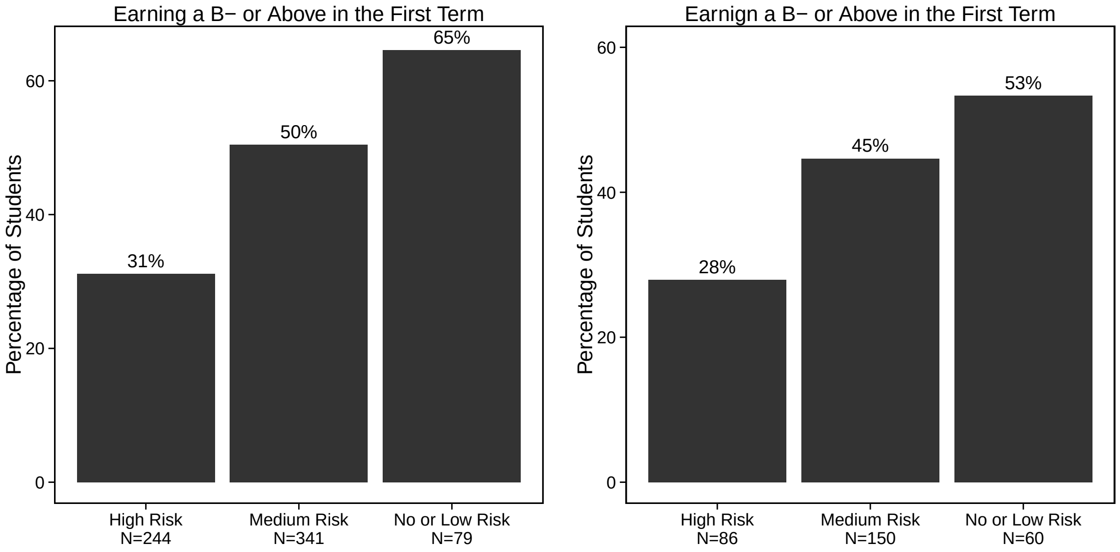
Figure 14: Cumulative risk and failing the common assessment. High risk: 4 through 9 risk factors; medium risk: 2 or 3 risk factors; no or low risk: 0 or 1 risk factors.



(a) Statway ($\chi^2(2)\, p = 0.021$)

(b) Quantway ($\chi^2(2)\, p = 0.001$)

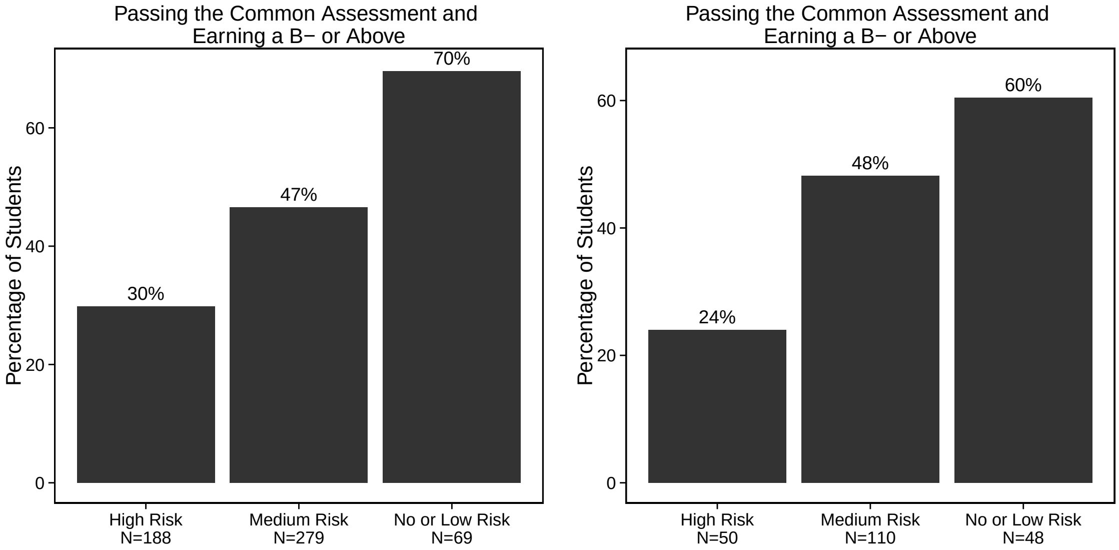
Figure 15: Cumulative risk and withdrawn from or failing to complete the Pathways. High risk: 4 through 9 risk factors; medium risk: 2 or 3 risk factors; no or low risk: 0 or 1 risk factors.



(a) Statway ($\chi^2(2)\, p < 0.001$)

(b) Quantway ($\chi^2(2)\, p = 0.005$)

Figure 16: Cumulative risk and earning a B- or above. High risk: 4 through 9 risk factors; medium risk: 2 or 3 risk factors; no or low risk: 0 or 1 risk factors.



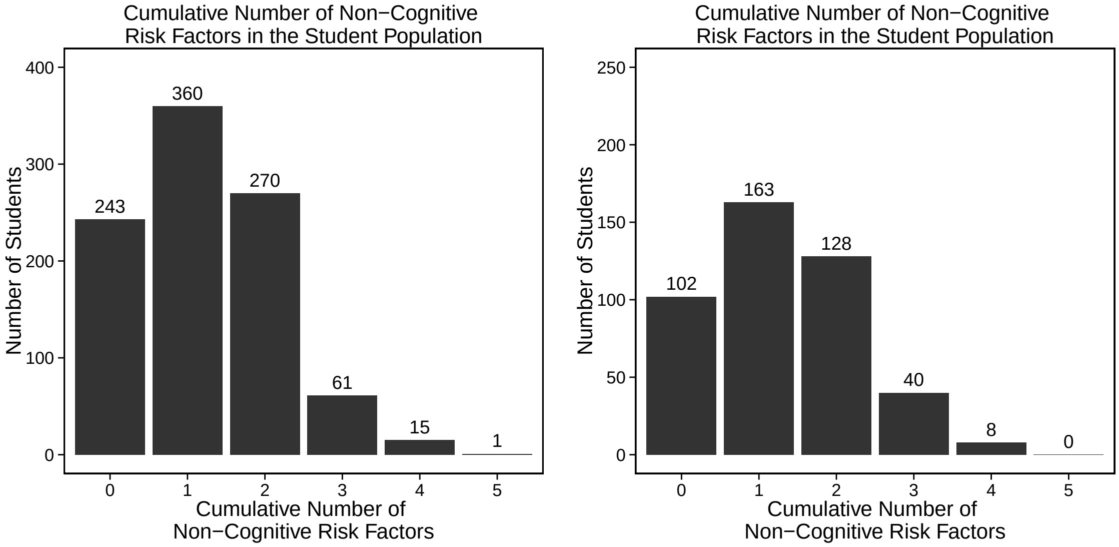
(a) Statway ($\chi^2(2)\, p < 0.001$)

(b) Quantway ($\chi^2(2)\, p = 0.001$)

Figure 17: Cumulative risk and passing the common assessment as well as earning a B- or above. High risk: 4 through 9 risk factors; medium risk: 2 or 3 risk factors; no or low risk: 0 or 1 risk factors.

Cumulative Productive Persistence Risk

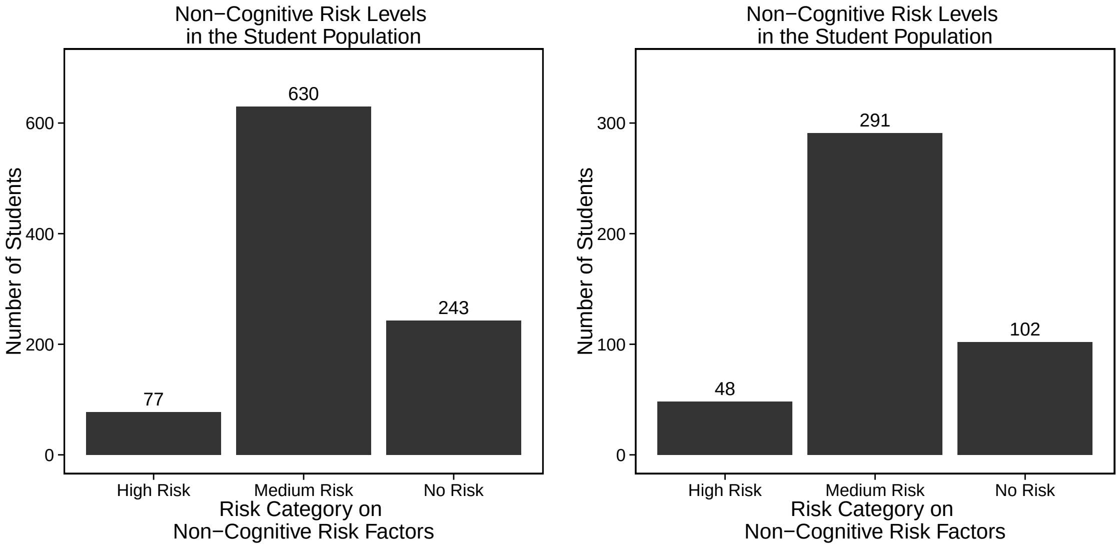
Among the nine "risk factors" are five productive persistence risk factors used to create the cumulative "non-cognitive at-riskness" measures, which range from 0 to 5. As with the total "at-riskness" index, students' "non-cognitive" totals can be reasonably grouped into three risk categories: low, medium, and high productive persistence risk. The following figures display the number of students in these "non-cognitive" risk categories and how the categories relate to different course outcomes.



(a) Statway

(b) Quantway

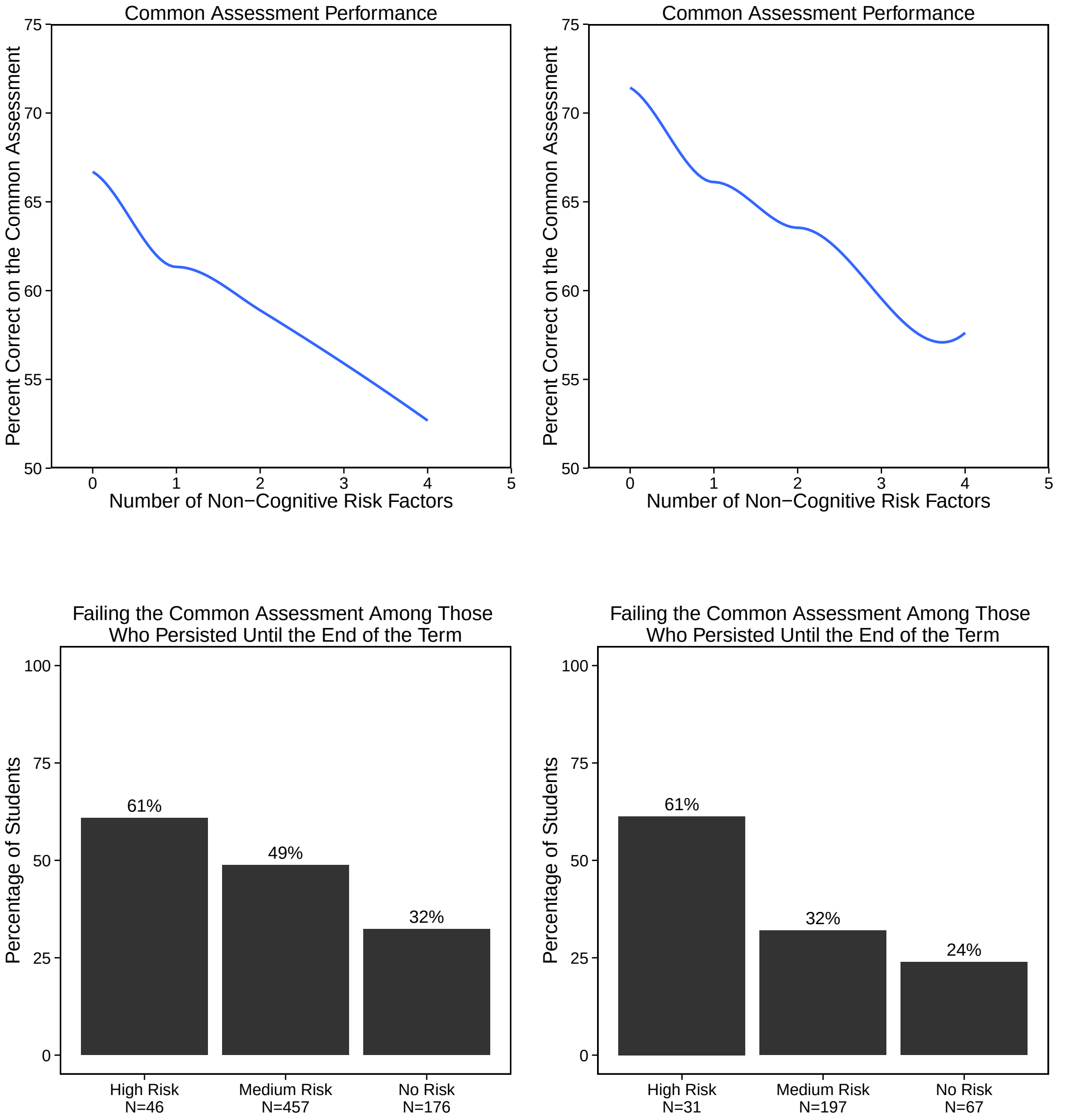
Figure 18: Cumulative number of non-cognitive risk factors.



(a) Statway

(b) Quantway

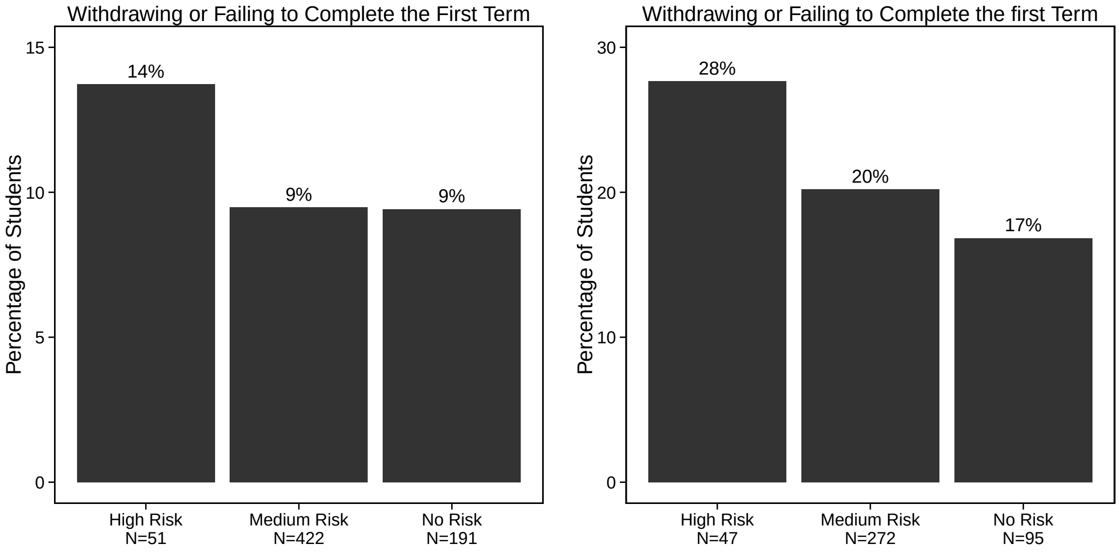
Figure 19: Cumulative non-cognitive risk category. High risk: 3 through 5 risk factors; medium risk: 1 or 2 risk factors; no risk: 0 risk factors.



(a) Statway ($\chi^2(2)\, p < 0.001$)

(b) Quantway ($\chi^2(2)\, p = 0.001$)

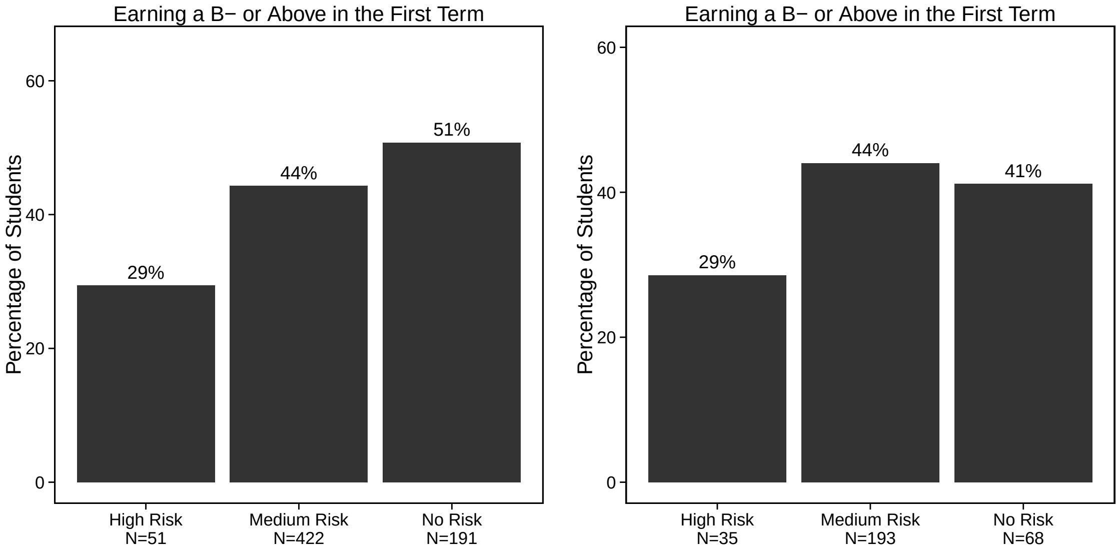
Figure 20: Cumulative non-cognitive risk and failing the common assessment. High risk: 3 through 5 risk factors; medium risk: 1 or 2 risk factors; no risk: 0 risk factors.



(a) Statway ($\chi^2(2)\, p = 0.616$)

(b) Quantway ($\chi^2(2)\, p = 0.32$)

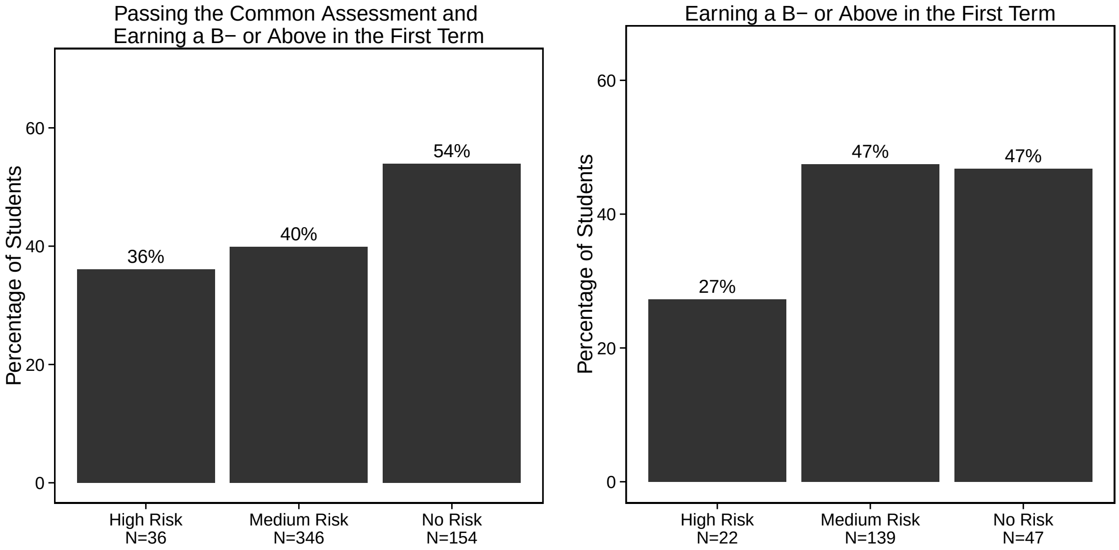
Figure 21: Cumulative non-cognitive risk and withdrawing from or failing to complete the Pathways. High risk: 3 through 5 risk factors; medium risk: 1 or 2 risk factors; no risk: 0 risk factors.



(a) Statway ($\chi^2(2)\, p = 0.022$)

(b) Quantway ($\chi^2(2)\, p = 0.232$)

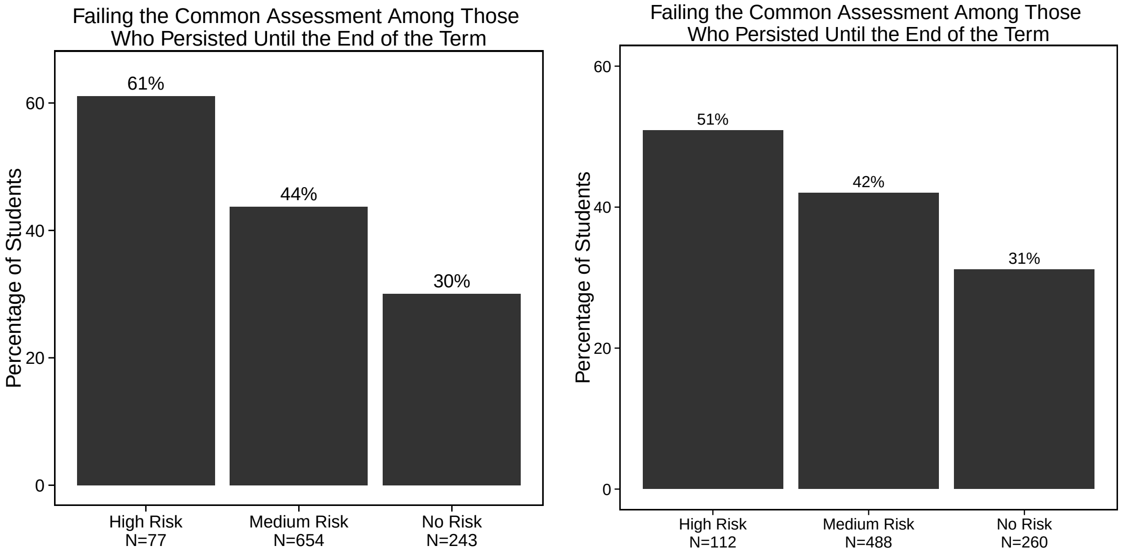
Figure 22: Cumulative non-cognitive risk and earning a B- or above. High risk: 3 through 5 risk factors; medium risk: 1 or 2 risk factors; no risk: 0 risk factors.



(a) Statway ($\chi^2(2)\, p = 0.009$)

(b) Quantway ($\chi^2(2)\, p = 0.202$)

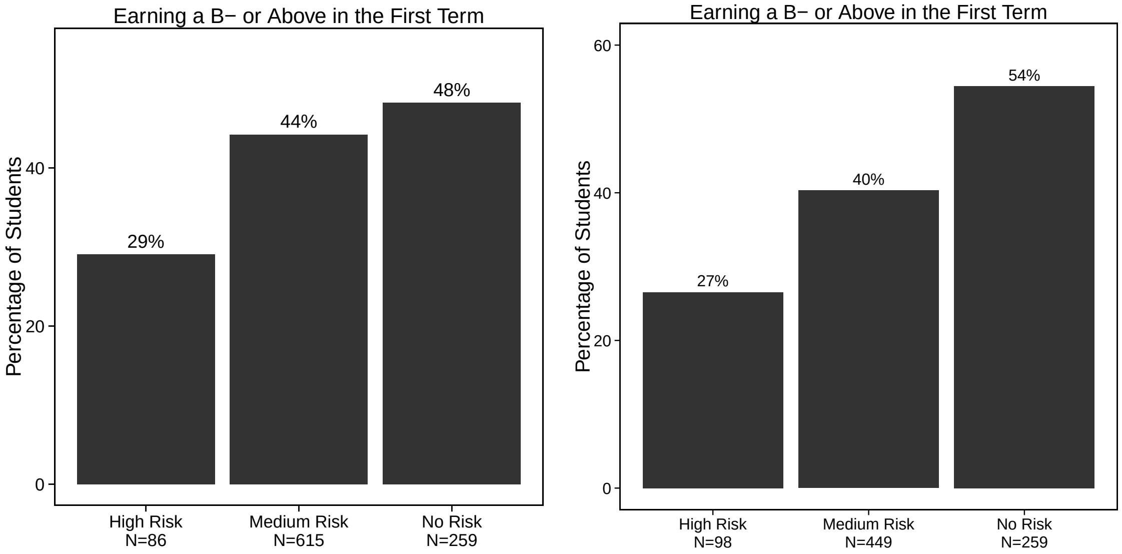
Figure 23: Cumulative non-cognitive risk and passing the common assessment as well as earning a B- or above. High risk: 3 through 5 risk factors; medium risk: 1 or 2 risk factors; no risk: 0 risk factors.



(a) 2011 ($\chi^2(2)\, p < 0.001$)

(b) 2012 ($\chi^2(2)\, p = .001$)

Figure 24: Cumulative non-cognitive risk and failing the common assessment in the Pathways. High risk: $3, 4$ or 5 risk factors; medium risk: 1 or 2 risk factors; no risk: 0 risk factors.

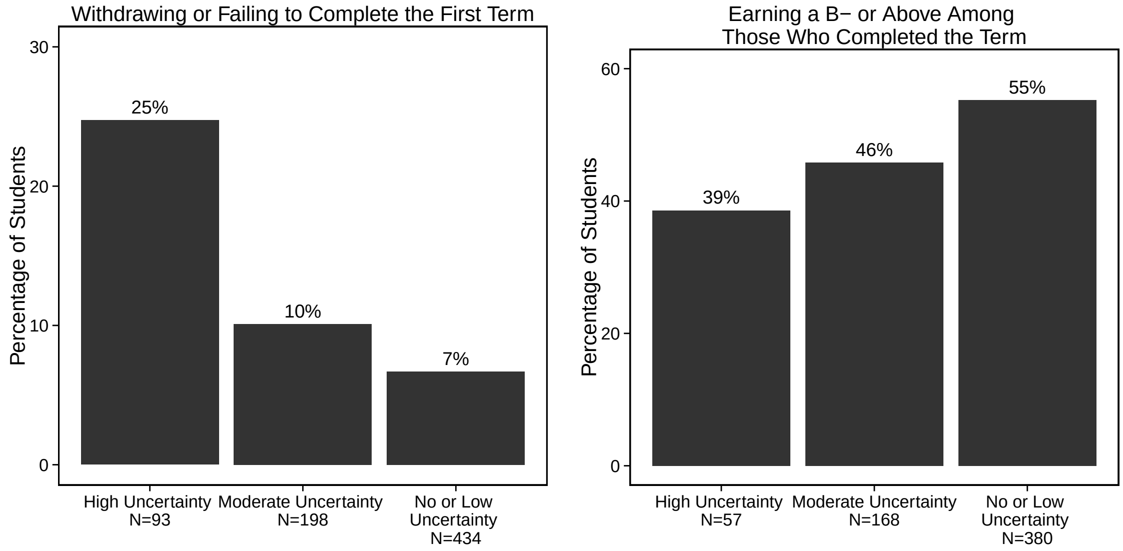


(a) 2011 ($\chi^2(2)\, p = 0.008$)

(b) 2012 ($\chi^2(2)\, p < .001$)

Figure 25: Cumulative non-cognitive risk and earning a B- or above in the Pathways. High risk: $3, 4$ or 5 risk factors; medium risk: 1 or 2 risk factors; no risk: 0 risk factors.

Analyzing Week Four Productive Persistence Data



(a) Course Withdrawal ($\chi^2(1)\, p < 0.001$)

(b) Earning a B- ($\chi^2(1)\, p = 0.018$)

Figure 26: Belonging uncertainty after four weeks. High uncertainty: response = 1 or 2; moderate uncertainty: response = 3; no or low uncertainty: response = 4 or 5.

Regression Analyses

Table 6: Number of productive persistence “at-riskness” factors on day 1 of the the Pathways courses predicts end-of-term performance on a common assessment. All independent variables are centered around the grand mean.

	b (se)
Intercept, γ_{000}	$-.05$ (.07)
Number of productive persistence “at-riskness” factors, γ_{100}	$-.09^{***}$ (.03)
Math conceptual knowledge,γ_{200}	$.35^{***}$ (.03)
Income, γ_{300}	.02 (.01)
African American, γ_{400}	$-.39^{***}$ (.07)
Hispanic/ Latino, γ_{500}	$-.11$ (.07)
Home language is English, γ_{600}	$-.19^{**}$ (.07)
Number of dependents, γ_{700}	.01 (.02)
Maternal Education, γ_{800}	.10 (.06)
Number of students	974
Number of courses	75
Number of colleges	28
Level 1 variance	.54
Level 2 variance	$.07^{***}$
Level 3 variance	$.10^{***}$
Note:	*p<0.1; **p<0.05; ***p<0.01.

Table 7: A one-item belonging uncertainty measure predicts semester-end course withdrawal and overall course grades in the Pathways. All independent variables are centered around the grand mean. Coefficients listed in the course withdrawal model are the unstandardized coefficients from a logistic regression model.

	Course Withdrawal (official or unofficial) b (se)	End-of-term grade (0=F, 4=A) b (se)
Intercept, γ_{000}	-2.32^{***}	2.38^{***}
	(.27)	(.10)
Belonging uncertainty, γ_{100}	$-.36^{***}$	$.14^{***}$
	(.11)	(.04)
Covariates		
Math conceptual knowledge, γ_{200}	$-.11$	$.26^{***}$
	(.13)	(0.04)
Income, γ_{300}	$-.06$	.01
	(.06)	(.02)
African American, γ_{400}	.15	$-.35^{**}$
	(.31)	(.12)
Hispanic/ Latino, γ_{500}	.24	$-.07$
	(.38)	(.13)
Home language is English, γ_{600}	.18	$-.02$
	(.36)	(.12)
Number of dependents, γ_{700}	.11	$-.01$
	(.08)	(.04)
Maternal education, γ_{800}	$-.12$	.18
	(.33)	(.11)
Number of students	725	605
Number of courses	69	60
Number of colleges	24	21
Level 1 variance	N/A	1.06
Level 2 variance	.08	$.07^{*}$
Level 3 variance	1.12^{***}	$.12^{***}$
Note:		$^{*}p<0.1$; $^{**}p<0.05$; $^{***}p<0.01$.