Thinkstock/iStock

Got grit? Maybe . . .

Self-reported survey data — concerning students' levels of grit or their mindsets — are all the rage. But beware of using such data for school improvement.

By Brent Duckor

Education historians have often documented the faddish nature of education reform. According to the late David Tyack and his colleague Larry Cuban (1990), schools in the United States are in the habit of reforming — over and over again.

Remember the aptitude tests in the 1940s? The vocational counseling agenda in the 1950s and 1960s? The personality testing and dispositions push in the 1970s? These reforms have come and gone, only to be resurrected years later under the guise of a new, improved solution for public education.

Today's student dispositions movement is such a trend. It's called by many names: socio-emotional learning outcomes, noncognitive indicators, affective factors, behavioral objectives and skills. The constructs that currently animate it are "grit" and "growth mindset." Grit-oriented reformers tell us

BRENT DUCKOR (brent.duckor@sjsu.edu) is an associate professor in the College of Education, San Jose State University, San Jose, Calif.

we can overcome the academic achievement gap by boosting nonacademic or noncognitive factors. Betting that noncognitive dispositions will make a difference in education outcomes, some policymakers are now taking their cues from Aesop's tortoise: Although the tortoise doesn't seem to possess the skills required to win the race, he does have the "grit" that enables him to prevail.

The problem with the dispositions bandwagon is that there is thin evidence for the reliability and instructional uses of noncognitive factors in K-12 schools. Even if grit does make a difference, it remains unclear how to assess that difference or to teach students how to be more "gritty." So the question is, what value do these noncognitive indicators have if they cannot provide reliable guides to improved teaching and learning? Educators aren't interested in psychological traits and factors in the abstract. Instead, our necessary focus is on the teaching and learning that happen daily in the classroom as students interact with conceptually difficult subject matter.

It is time to ground the dispositions discourse in the principles of educational assessment.

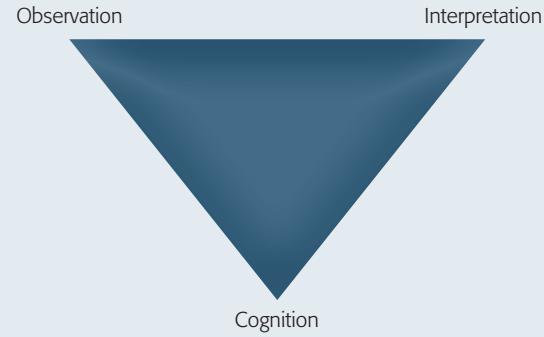
The assessment triangle

We've known for decades that psychological factors are not the same as educational measures. Unlike psychological researchers who work with relatively small, voluntary samples of subjects, educational assessment experts work in high-stakes settings, with significant consequences for children in public schools, many of whom represent vulnerable, historically disadvantaged groups.

If grit is the x factor that drives an accountability index or educational policy push, it surely will be hard to observe without bias.

Educational assessment experts have developed the Standards for Educational and Psychological Testing (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 2014). These are validation guidelines for reviewing fairness and bias in addition to what experts call "a logic of assessment," to guard against overgeneralization, spurious claims, and invalid inferences from data. Let's take a moment to review the logic of educational assessment and see how it's represented in the Assessment Triangle, a framework for understanding the connections among what students know, how we might observe their performances, and how we might know if they've acquired knowledge and skills in a meaningful way (see Figure 1).

FIGURE 1.
The assessment triangle



Using the Assessment Triangle as a guide, the National Research Council (Pellegrino at al., 2001) noted that every assessment is based on three interconnected principles:

> A theory of what students know and how they develop competence in a subject domain (cognition), tasks or situations used to collect evidence about student performance (observation), and a method for drawing inferences from those observations (interpretation) (p. 36).

Principle #1: Cognition

Key to the logic of assessment is the notion that everything in K-12 education rests on how students learn and master a subject. Measurement specialists and test designers must carefully attend to student cognition, the bottom-most vertex of the triangle. In the noncognitive domain, this would mean having a theory of how students actually progress in acquiring and demonstrating a disposition, for example, grit.

We know that the logic of assessment stands (or falls) on how well assessment specialists articulate how students develop competence in a particular context, such as mathematics, physical education, or history. Educational assessment experts want to capture trajectories, linking formative and summative indicators, and modeling how learners change in developmentally sensitive ways.

However, utterly absent from the dispositions discourse is a basic theory of growth and student change over time. Research has offered examples of learning progressions, facets of knowledge and skills, and a host of other taxonomical reference points (Bransford et al., 1999; Pellegrino et al., 2001) that guide the question, What are we assessing? Teachers and

assessment experts know that choosing meaningful learning targets tied to student cognition is half the battle in terms of assessing students well, particularly if one is to discover what they know and can do in a given curriculum or grade level.

If dispositions advocates can't identify clear learning targets or a theory of student noncognition, then they haven't met the basic test of the logic of assessment. There is no excuse for ignoring the bottom vertex of the Assessment Triangle, the first principle of sound educational assessment.

Adding new labels to reinforce stereotypes about low-income student dispositions may be even more pernicious than failing to expose the socioeconomic roots of the academic achievement gap.

Principle #2: Observation

This principle reminds us that assessment tools should be appropriately aligned with a well-defined set of learning outcomes rooted in the first vertex, cognition and student growth. But noncognitive research relies almost exclusively on data derived from Likert surveys instead of drawing from a variety of instruments and robust item design (Wilson, 2005). Educational psychologists working in the noncognitive field seem to favor this item format, whether or not it supports the first principle in the Assessment Triangle.

Likert scales in the noncognitive realm present children with sets of statements or items to choose from, typically ranging from "strongly disagree" to "strongly agree." However:

- Can children really "progress" on these artificially manufactured five-point, six-point, or seven-point "scales," assuming they understand the wording and answer "correctly"?
- Do Likert-style item formats used to elicit responses represent any underlying theory of student cognition or change, for example, based on the work of Piaget, Vygotsky, or today's cognitive learning theorists?
- How do we know these Likert items best represent variation in our students' experience, such as levels or dimensions on a continuum of learning and teaching, and not just tallies of forced survey choices?

There are well-documented problems with using self-report surveys to evaluate complex psychological constructs, just as there is concern about using multiple-choice test formats to assess higher-order thinking skills and performances in schools that serve vulnerable, historically disadvantaged students (Darling-Hammond et al., 2015; Duckor & Perlstein, 2014; Resnick & Resnick, 1992). Dumbing down a complex construct by applying a one-size-fits-all survey does not lead to educational progress for students.

Principle #3: Interpretation

The interpretation vertex is perhaps the most important element of the Assessment Triangle as it addresses the fundamental challenge of making valid interpretations of data, noncognitive or otherwise. However, the literature on evaluating students' dispositions using Likert surveys gives little attention to this part of the Triangle, and it does little to establish the types of validity and reliability evidence needed to make a credible argument for the use of such surveys in K-12 classrooms, whether the data is to be used for diagnostic, formative, or summative purposes. (Nonetheless, the creators of noncognitive survey tools generally neglect to include a fair use clause or buyer beware notice.)

Further, while validity is necessary, it is not sufficient to make claims about the strength of students' noncognitive dispositions. Reliability issues abound, too, particularly as schools and districts borrow and modify noncognitive survey items to feed into their accountability platforms. All else being equal, the fewer items on these disposition surveys, the less reliable they will be for helping educators make sound judgments based on student scores. There is simply too much extraneous noise surrounding these Likert survey results including fatigue, faking, and what even advocates call reference bias (West, 2016).

Join the conversation

facebook.com/pdkintl
@pdkintl

Key to the logic of assessment is the notion that everything in K-12 education rests on how students learn and master a subject.

Advocates of the dispositions discourse have not addressed the basic test of the logic of assessment when they don't identify clear learning targets or a theory of student noncognition.

Another problem with the interpretation of noncognitive "scales" is whether surveys really are scales in the strict sense. Are there really "levels of grit" on a standard unit of measurement, or are there just varied voting results based on predetermined categories imposed on students? We know that too many of these so-called psychological scales are, in fact, sample- and question-dependent and therefore yield unstable results (Embretson & Reise, 2000).

The National Research Council (Pellegrino et al., 2001) emphasizes that it is "crucial [that] each of the three elements [cognition, observation, and interpretation] . . . not only must make sense on its own but also must connect to each of the other elements in a meaningful way to lead to an effective assessment and sound inferences" (p. 49). In evaluating the claims made about students by noncognitive researchers, we need to look much more closely at the quality of their inferences and the evidence trail provided by their instruments to support appropriate, intended uses.

The bottom line: Noncognitive "data" and "metrics" used by reformers to support decisions to group and match schools by levels of grit, or to place students in particular interventions to improve mindset, must be validated with common sense and a body of technical evidence (see Kane, 2013). The first principles of the science and design of assessment apply not only to traditional, summative achievement testing but also to noncognitive survey instruments as they scale up and begin to serve as new accountability indicators.

Remembering the lessons we've learned

Perhaps we all have been acculturated into believing that Likert surveys get the job done. More troubling is the notion that these noncognitive surveys all pretty much measure the same thing. When educational policymakers come to believe that all noncognitive surveys are essentially interchangeable, we must pause. Are we repeating the troubling history of educational testing rather than learning from it?

As Popham (2004) presciently wrote over a decade ago, addressing the misguided belief that a test is a test is a test:

> This misperception has serious consequences. For example, every few weeks we are apt to find newspaper reports describing investigations whose results have clear implications for education policymakers. Such studies might contrast the successes of public schools versus private schools, charter schools versus noncharter schools, and board-certified teachers versus non-board-certified teachers. However, before placing confidence in such empirical investigations — especially in studies that may influence the way we educate our children — we need to be certain the researchers adhered to the fundamental canons of research design (p. 88).

As we attempt to measure students' noncognitive skills in this era of accountability and increased public scrutiny of public schools, we can expect the media to seize these "numbers" and make hay over which students are "grittier" than others. Reporters, bolstered by economists and real estate agents, will be more than pleased to calculate the level of noncognitive achievement — by school, district, and state. Working with a misconception similar to the one Popham (2007) addressed in standardized testing circles, these watchdogs will gladly assume that a given student's "grit" is exemplary or deficient, no matter what sorts of surveys are used. In most instances, they won't even supply the names of the surveys or the technical evidence to support particular purposes and uses.

Instead, the media is more likely to succumb to the misguided notion, as they have with educational achievement indicators, that a survey is a survey is a survey. That is, they will regard even substantively different self-report surveys as essentially interchangeable. In most instances, this is simply not true.

Many of these noncognitive surveys purport to measure a child's "character" without actually calling it that. Moral reformers in the United States have long sought to bring virtue to schools. Adding new labels to reinforce stereotypes about low-income student "dispositions" may be even more pernicious than failing to expose the socioeconomic roots of the academic achievement gap.

It matters — but can we measure it?

Whether one focuses on perseverance, social skills, academic mindset, learning strategies, or behaviors and skills (Farrington et al., 2012), we must confront the challenges ahead. It's time to address noncognitive research and the "multiple measures" that are most likely to lead to unintended consequences in K-12 schools.

When it comes to assessing student attitudes and beliefs, teachers will undoubtedly see the challenges differently than will school administrators or district accountability directors. Charged with acting on "the data," teachers may wonder, Who will report to us on our students' level of grit? What will we be expected to do with this information? How instructionally sensitive are these noncognitive surveys, and what do they tell us about the intersection of dispositions with our students' content acquisition? Do my students exhibit more grit in their study of art and music than they do in math or science? How will we benchmark progress with "growth in grit," for example?

Parents also will have a different take on what having "low" or "high" levels of grit or academic "mindset" means for their children. They might want to know: Should I move my child to "grit-centric" programs and look for the "grittiest" teachers the school has to offer? Can schools and districts be held accountable for zero growth on grit? Should lawyers and civil rights advocates add grit and academic mindset to the school conditions and resources list in equity cases?

Dumbing down a complex construct by applying a one-size-fits-all survey does not lead to educational progress for students.

And what is to be done with the think tanks, economists, and statisticians? Soon someone will run a quantitative analysis or value-added investigation, only to discover that the results indicate that there is no significant difference between the noncognitive achievement of students from one group and the achievement of students from another group. Should they (or we) be outraged by this finding?

Again, Popham (2004) puts the point more sharply in the achievement domain, but the same logic holds for the abuses of noncognitive tests and self-reported survey data:

> I want to shriek out, "On which tests?" If the achievement tests being used are strongly influenced by socioeconomic status, then there's really no point in carrying out a study that uses those instructionally insensitive tests to measure the effects of instruction. Unless reporters describe the specific achievement tests being used — so that interested readers can at least consider the likely instructional sensitivity of those tests — it is folly to place any real confidence in a study's conclusions (p. 88).

In the world of educational achievement, we're focused on reading, writing, and other academic domains that are fairly concrete. You can read a text and try to decipher its meaning. You can write an essay and try to make a persuasive point. You can even dribble a basketball, play a piece of music, or design a robot. There's a tangible object or process that "we" can observe in a discrete amount of time, as educational assessment experts and test makers.

The same is not true for noncognitive factors. If grit is the x factor that drives an accountability index or educational policy push, it surely will be hard to observe without bias. As Rees (2013) notes,

> MacArthur Genius grantee and University of Pennsylvania psychologist Angela Duckworth is one of the foremost experts on this topic and has also had remarkable success making the concept palatable to the media, perhaps in part thanks to a blunt distillation: She simply calls it "grit." Neither too cerebral nor too soft nor too corporate, it's easy to see the appeal. But is Duckworth's definition — "the tendency to sustain interest in and effort toward very long-term goals" — inclusive enough? Does it allow for all the complexity and nuance of the interplay among attitudes, beliefs, skills, and behaviors that we want to see in a robust approach to noncognitive skills? Does emphasizing the importance of holding on tightly to a singular goal set us up to put less value on flexibility, adaptability, or innovation?

Confusion over the meaning of student disposition and how best to study it obscures a deeper tension in what educational assessment experts call *construct validity* (the magical "it" in grit, if you will). We just don't know how to validate the claims and proposed uses of this sort of data for school children.

Join the conversation

facebook.com/pdkintl
@pdkintl

Be wary

No amount of hand-waving about the value of multiple measures in education reform can hide the obvious. More research needs to be done — and more careful attention to the logic of assessment is required — before we start adopting this or that noncognitive survey.

If it turns out under ESSA and new state-led approaches to accountability that self-report surveys used to evaluate student dispositions are instructionally insensitive, then we are back to square one.

While researchers in the noncognitive domain increasingly have come to recognize the challenges with ascribing scores to K-12 students (Duckworth & Yeager, 2015), there has been less attention to the assessment design principles (Pellegrino et al., 2001) and testing standards (AERA, APA, & NCME, 2014) that might better guide the policymakers poised to adopt new indicators. Already we see states grappling to build and borrow school climate and conditions indicators. Tight budgets,

limited attention, and shifting policy priorities will lead to borrowing off the shelf surveys. When the primary interest is aggregating the results to support a finding about educational achievement, shortcuts are inevitable.

Cuban and Tyack remind us: With the new knowledge economy — and its ceaseless demand for low-cost solutions and marketable educational technologies — policymakers and vendors can accelerate a fad, are likely to make it a trend, and turn it almost overnight into a nationwide movement.

We need to stop and pause. If we continue down the dispositions path, it won't be long before schools — and therefore teachers and students — are handed a survey solution that sets back the clock on what we've known for decades about the logic of assessment and best test design. K

References

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2014). *Standards for educational and psychological testing.* Washington, DC: Author.

Bransford, J., Brown, A.L., Cocking, R.R., & National Research Council. (1999). *How people learn: Brain, mind, experience, and school.* Washington, DC: National Academies Press.

Cuban, L. (1990). Reforming again, again, and again. *Educational Researcher, 19* (1), 3-13.

"I would solve this story problem by developing the characters, heightening the conflict, and concluding with a satisfactory resolution."

Darling-Hammond, L., Abedi, J., Adamson, F., Chingos, J., Conley, D.T., Falk, B., et al. (2015). *Beyond the bubble test in next generation assessment.* San Francisco, CA: John Wiley & Sons.

Duckor, B. & Perlstein, D. (2014). Assessing habits of mind: Teaching to the test at Central Park East Secondary School. *Teachers College Record, 116* (2), 1-33.

Duckworth, A.L. & Yeager, D.S. (2015). Measurement matters: Assessing personal qualities other than cognitive ability for educational purposes. *Educational Researcher, 44* (4), 237-251.

Embretson, S.E. & Reise, S.P. (2000). *Item response theory for psychologists.* Mahwah, NJ: Lawrence Erlbaum Associates.

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

Kane, M. (2013). Validation as a pragmatic, scientific activity. *Journal of Educational Measurement, 50* (1), 115-122.

Pellegrino, J.W., Chudowsky, N., Glaser, R., & National Research Council. (2001). *Knowing what students know: The science and design of educational assessment.* Washington, DC: National Academies Press.

Popham, W.J. (2004). A test is a test is a test — Not. *Educational Leadership, 64* (4), 88-89.

Popham, W.J. (2007). Instructional insensitivity of tests: Accountability's dire drawback. *Phi Delta Kappan, 89* (2), 146-150.

Rees, N. (2013, October 21). Grit: The new education flavor of the month? *U.S. News & World Report.* www.usnews.com/opinion/blogs/nina-rees/2013/10/21/macarthur-genius-angela-duckworth-grit-and-education-reform

Resnick, L.B. & Resnick, D.P. (1992). Assessing the thinking curriculum: New tools for educational reform. In B.R. Gifford & M.C. Connor (Eds.), *Changing assessments: Alternative views of aptitude, achievement, and instruction* (pp. 37-75). Boston, MA: Kluwer Academic.

West, M.R. (2016). Should noncognitive skills be included in school accountability systems? Preliminary evidence from California's Core districts. Brookings. *Evidence Speaks Reports, 1* (13). www.brookings.edu/research/reports/2016/03/17-non-cognitive-skills-school-accountability-california-core-west

Wilson, M.R. (2005). *Constructing measures: An item response theory approach.* Mahwah, NJ: Lawrence Erlbaum Associates.