Current Frontiers of Non-Cognitive Measurement: Insights for Policy and Practice

BY LISA QUAY

RESEARCH BRIEF | MAY 2015

This Research Brief summarizes the article by Angela Duckworth and David Yeager in *Educational Researcher*, "Measurement Matters: Assessing Personal Qualities Other Than Cognitive Ability for Educational Purposes." Based on the authors' analysis of the field, we offer recommendations for how multiple stakeholders can advance the state of non-cognitive measurement for practical purposes in education.

Overview

In recent years, education practitioners and policymakers have become increasingly interested in so-called "non-cognitive" qualities that research has shown are predictive of educational and life outcomes.[i] These qualities range from behaviors (e.g., self-control) and beliefs (e.g., about the nature of intelligence) to skills (e.g., interpersonal communication and conflict resolution).

The burgeoning interest in these qualities has been accompanied by a desire to measure them to improve student outcomes. And there is over 40 years of scholarly research on non-cognitive measures to draw on. In this brief, we describe the measures that have emerged from this research, present their strengths and weaknesses for applied use, and offer recommendations for advancing the field of non-cognitive measurement for educational purposes.

We explain that existing questionnaire measures are particularly ill-suited for accountability purposes and should not be used to rate individual students' non-cognitive qualities for purposes of diagnosis. If properly adapted and refined, questionnaires and performance tasks may hold great promise for use in program evaluation and practice improvement under some circumstances.

We recommend that practitioners, researchers, program evaluators, and developers collaborate on five key tasks that could help unlock the potential of non-cognitive measures for educational purposes:

1. Optimizing existing questionnaires for purposes of practice improvement;
2. Conducting R&D to generate better performance tasks for program evaluation;
3. Integrating seamless measurement and reporting in popular online education platforms;
4. Building an online non-cognitive measurement repository and reporting tool for educators; and,
5. Increasing educators' facility with practical measurement and the interpretation of data on non-cognitive measures so that educators can use this information to improve classroom practice.

MINDSET SCHOLARS NETWORK

Hosted at the Center for Advanced Study in the Behavioral Sciences at Stanford University, the Mindset Scholars Network is a group of leading social scientists dedicated to improving student outcomes and expanding educational opportunity by advancing our scientific understanding of students' mindsets about learning and school.

Types Of Non-Cognitive Measures

Broadly speaking, there are two categories of measures that researchers have designed to capture students' non-cognitive qualities: questionnaires and performance tasks.

Questionnaires

Researchers rely on two types of questionnaires: (a) **self-report questionnaires completed by students,** and (b) **questionnaires completed by adults—typically teachers—about students.** (See inset for sample questionnaire items.) Questionnaires are popular because they are cheap, quick to administer, produce consistent results, and often predictive of objective outcomes.[ii]

Nevertheless, questionnaires are susceptible to problems that diminish the degree to which they actually measure what they claim to measure. For example, students in schools in which there are strict behavior norms tend to rate their own self-control more stringently than their peers completing the same questionnaire in schools with relatively lax standards—even when the former group's measured behavior may suggest greater self-control in reality (an issue known as "reference bias").[iii] Table 1 describes five such potential challenges in using questionnaires.

Sample Student Self-Report Non-Cognitive Questionnaire Items

GROWTH MINDSET: "You have a certain amount of intelligence and you really can't do much to change it."*

☐ Strongly Agree ☐ Agree ☐ Mostly Agree
☐ Mostly Disagree ☐ Disagree ☐ Strongly Disagree

GRIT: "I have achieved a goal that took years of work."

☐ Very much like me ☐ Mostly like me
☐ Somewhat like me ☐ Not much like me
☐ Not like me at all

Note: *The extent to which students disagree with fixed mindset statements like this is how researchers measure growth mindset beliefs.

SOURCE: Blackwell, L. S., et al. (2007). Implicit theories of intelligence predict achievement across an adolescent transition: A longitudinal study and an intervention. *Child Development, 78,* 246-263. Duckworth, A. L., et al. (2007). Grit: Perseverance and passion for long-term goals. *Journal of Personality and Social Psychology, 92,* 1087-1101.

Sample Non-Cognitive Performance Task

The most famous example of a performance task in psychology is Dr. Walter Mischel's "Marshmallow Test", which assesses delayed gratification. In this task, preschoolers are presented with a smaller pile of treats and a larger pile of treats. The children are then told that they can have the smaller pile now, or they can receive the larger pile if they wait for the experimenter to come back in the room. The amount of time children can wait has been found to predict later life outcomes.

SOURCE: Mischel, W. (2014). *The Marshmallow Test: Mastering self-control.* New York, NY: Little, Brown.

Table 1. Potential Weaknesses of Student Self-Report and Teacher-Report Questionnaires

POSSIBLE THREATS TO VALIDITY	EXPLANATION
1. Misinterpretation by participant	Student or teacher may read or interpret the item in a way that differs from the researcher's intent
2. Lack of insight or information	Student or teacher may not be astute or accurate reporters of behaviors or internal states (e.g., emotions, motivation) for a variety of reasons
3. Insensitivity to short-term changes	Questionnaire scores may not reflect subtle changes in the non-cognitive quality over short periods of time
4. Reference bias	The frame of reference (i.e., implicit standards) used when making judgments may differ across students or teachers
5. Faking and social desirability bias	Students or teachers may provide answers that are desirable but not accurate

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

Performance Tasks

An alternative to questionnaires is what researchers call "performance tasks". These tasks are **situations that have been carefully constructed to allow researchers to observe meaningful differences in certain behaviors.** (See inset, previous page, for an example of a performance task.)

A primary advantage of performance tasks is that they are not beholden to the subjective judgments and reports of students or teachers. In this way, they avoid the reference bias, social desirability bias, acquiescence bias (a tendency to agree with statements), and faking that plague questionnaires. Additionally, since they assess a behavior at a particular point in time, they can be more sensitive to subtle changes in behavior that may not be picked up on a questionnaire.

Performance tasks, however, have their own set of drawbacks (see Table 2). Objective tasks still require researchers to make subjective inferences about the internal motivations, feelings, and thoughts of the participant. (For example, when a child plays with his toys when instructed to do so, is he demonstrating autonomous self-control or compliance with adult authority?) The validity of performance tasks is also threatened by repeated exposure to the same task; while this can be addressed through a battery of performance measures, this solution can be costly in terms of time and expense.[iv] Moreover, such tasks are dependent on ensuring carefully controlled conditions, which can be challenging and costly, and difficult to achieve since situational factors that affect task performance can vary systematically across groups.

Table 2. Important Limitations of Performance Tasks

POSSIBLE THREATS TO VALIDITY	EXPLANATION
1. Misinterpretation by researcher	Researchers may make inaccurate assumptions about underlying reasons for student behavior
2. Insensitivity to typical behavior	Tasks which optimize motivation to perform well (i.e., elicit maximal performance) may not reflect behavior in everyday situations
3. Task impurity	Task performance may be influenced by irrelevant competencies (e.g., hand-eye coordination)
4. Artificial situations	Performance tasks may foist students into situations (e.g., doing academic work with distracting videogames in view) that they might proactively avoid in real life
5. Practice effects	Scores on sequential administrations may be less accurate (e.g., because of increased familiarity with the task or boredom)
6. Extraneous situational influences	Task performance may be influenced by aspects of the environment in which the task is performed or by the physiological state (e.g., time of day, classroom noise, hunger)
7. Random error	Scores may be influenced by purely random error (e.g., respondent randomly making choices on the task)

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

Issues To Bear In Mind When Considering Non-Cognitive Measures For Specific Educational Purposes

The validity of a measure is not inherent to the measure itself, but rather is a characteristic of a measure with regard to a particular use.[v] **When considering the use of a measure, one must assess both the properties of the measure and its intended application.**

It is important to remember that most non-cognitive measures were designed and validated for academic research. Measures designed for research have unique attributes: they need only be used once or twice with an individual student, and the results are aggregated across hundreds or even thousands of students to understand the relationship between a non-cognitive quality and longer-term outcomes. The nature of measurement for educational purposes is in many cases quite different. For this reason, any application of these measures outside of research warrants careful consideration and caution.

There are four primary applications for which there is demand for measures of non-cognitive qualities: (a) program evaluation; (b) accountability; (c) individual diagnosis; and (d) practice improvement. Below we describe which measures are best suited—and ill-advised—for these various educational purposes.

A. Program Evaluation

Evaluations document how a program is being implemented and whether it is having its intended effect, and are thus essential resources for both practice and policy. As illustrated in the reference bias example above, student self-report measures may be particularly problematic for examining *within-person program effects* (i.e., assessing a change between a pre- and post-test) or *between-program differences* (i.e., average differences between schools or programs).[vi] While teacher-report questionnaire measures may be valid for program evaluation that is restricted to *within a single school,* using them for between-school program evaluations (a far more common practice) can be problematic due to variation in frames of reference across schools, which can give rise to bias.

Performance tasks may be a viable alternative to questionnaires for program evaluations. They measure objective, quantifiable behaviors and are not subject to reference bias over time or across school sites. As noted above, however, current performance tasks are not without their own limitations. We believe a medium-term solution could be to create a suite of performance tasks that are both brief and scalable, as well as age-specific, and that could be easily administered in a group setting. This would almost certainly entail a set of online tasks accompanied by easy-to-follow administration protocols. **If practice effects could be reduced, a suite of performance tasks could be invaluable to program evaluation.**

B. Accountability

In contemporary K-12 education policy, the most utilized tool for encouraging desired behaviors and outcomes is the application of standardized measures in an accountability system. While existing accountability systems feature cognitive measures, there is growing interest in including non-cognitive measures, as well. The inherent limitations of questionnaires measuring non-cognitive qualities suggest they are particularly ill-advised for this use—regardless of whether the stakes are high (i.e., bearing the threat of consequences) or low (i.e., lower performers are required to learn from higher performers).

A primary concern with both questionnaire measures and performance tasks for accountability is the possibility of faking or inappropriately manipulating data. Another concern regarding questionnaires is the problem of reference bias, whereby students and educators who share more stringent expectations for behavior, for example, provide harsher ratings on questionnaires. This poses a serious challenge, particularly for accountability judgments between schools. Additionally, similar to value-added methods, aggregated student self-report data may only be useful for detecting schools at the very top and bottom of the distribution, but may be unable to distinguish between schools who are closer to average. One exception may be the use of such measures for within-school comparisons, which could help

The validity of a measure is not inherent to the measure itself, but rather is a characteristic of a measure with regard to a particular use. When considering the use of a measure, one must assess both the properties of the measure and its intended application.

The inherent limitations of questionnaires measuring non-cognitive qualities suggest they are particularly ill-advised for [accountability]—regardless of whether the stakes are high (i.e., bearing the threat of consequences) or low (i.e., lower performers are required to learn from higher performers).

schools identify "positive outlier" teachers who could serve as coaches to their peers. This, too, is similar to value-added measures using state test scores, which are more effective at distinguishing between teachers in the same school than between teachers in different schools.[vii]

Overall, we strongly caution that existing non-cognitive measures not be used for accountability purposes, regardless of how "low" or "high" the stakes, until improved measures are developed.

C. Individual Diagnosis

Schools often wish to assess individual students' cognitive capabilities for purposes of remediation or tracking. More recently, this has expanded to include an interest in rating individual students' non-cognitive qualities for such purposes. In this case, there are two major concerns with existing non-cognitive measures. The first regards reliability; it is unlikely that existing questionnaires would yield consistent ratings ofstudents' non-cognitive qualities for a given individual. Not even clinical depression scales can diagnose individuals using comprehensive questionnaires.[viii] A second issue involves how the assessment context can change students' performance on non-cognitive measures. For example, being worried about confirming negative stereotypes about their group (i.e., "stereotype threat") can make it hard for individuals to focus on the task at hand, or trust that they will be treated fairly.[ix] Assessments made under conditions that trigger stereotype threat could thus make certain groups appear to have lower levels of non-cognitive qualities than is true in reality.

In sum, we do not recommend that existing student self-report or teacher-report measures be used for individual diagnostic purposes. Moreover, even if more sophisticated protocols were developed, it would be critical that potential situational and group-specific biases be considered.

D. Practice Improvement

Professional educators are deeply concerned with continuous improvement of practice. "Practical measurement" plays an essential role in such efforts.[x] To be effective, practical measures must be able to be woven into daily instruction, sensitive to short-term changes, and quickly reported and readily analyzed by educators. Current questionnaires meet few of these requirements: they can be lengthy and are rarely customized for different settings. Moreover, they are not designed to be sensitive to short-term changes. That said, our experience suggests this may be a solvable R&D problem.

Performance tasks may be particularly useful for practice improvement because they can capture within-person changes over a short period of time.

Performance tasks may be particularly useful for practice improvement because they can capture within-person changes over a short period of time. To the extent that they could be delivered online and modified to reduce systematic and random error, a next generation of performance tasks could provide timely feedback to teachers and inform their improvement efforts.

In sum, while existing non-cognitive measures have limitations for practice improvement, they also have great promise. R&D efforts along these lines could yield substantial returns.

Summary and General Recommendations

As with any measure, the instruments developed by researchers to quantify non-cognitive qualities have both limitations and advantages. These attributes are important to bear in mind when considering such measures for settings and purposes for which they were not designed. **Research suggests limitations of current non-cognitive measures often undermine their validity for educational purposes:**

- Current questionnaires used for between-school and within-school comparisons, or over-time comparisons, may produce the *opposite* finding of the truth, and are ill-advised for accountability
- Existing questionnaire measures and performance tasks are insufficiently reliable to use for individual diagnostic purposes, and may produce erroneous results for certain groups
- Questionnaires and performance tasks may be useful for program evaluation and practice improvement, provided certain challenges with existing measures are resolved through R&D

We recommend that practitioners, researchers, program evaluators, and developers work collaboratively on five key tasks that could help unlock the potential of non-cognitive measures for practice improvement and program evaluation:

1. Optimize questionnaires for practice improvement purposes. Researchers could work with instructional experts to modify existing questionnaires to make them shorter; customizable for different settings, disciplines, and age groups; and deliverable during the course of instruction.

2. Conduct R&D to generate better performance tasks for program evaluation. Experts in program evaluation and measurement could work together to create new performance tasks that meet the practical needs of evaluators and are validated for use in estimating within-person effects and average differences between programs or schools.

3. Integrate seamless measurement and reporting in popular online education platforms. Researchers and online educational providers might collaborate on (a) new measures based on data that can be easily collected while students are engaged in online learning activities (e.g., analysis of failure patterns) and (b) reporting mechanisms that rapidly share results aggregated at the classroom-level with educators. Ethical oversight for these practices would be crucial.

4. Build an online practical measurement repository and reporting tool. Consider developing an online data collection and reporting platform that features questionnaires and performance tasks on an array of non-cognitive qualities, which educators could use to improve practice.

5. Increase educators' facility with practical measurement and the interpretation of data on non-cognitive measures. Consider adding pre-service training and professional development for educators on practical measurement, what non-cognitive measures validated for practice can (and can't) tell them, and how they can integrate them in continuous improvement efforts in their classrooms. This could be an important complement to efforts currently underway to identify instructional practices that foster non-cognitive qualities.

There is great power in measurement. Without measurement, it is impossible to know whether the changes we make are moving us in the right direction. Along with this power, comes the potential for misuse. We believe the field of non-cognitive measurement is no exception. We urge practitioners, policymakers, and funders to bear in mind the particular limitations and advantages of the existing non-cognitive measures that have been developed for research purposes, and to invest in the R&D and training necessary to yield measures and measurement practices that could empower those seeking to cultivate these important qualities in students.

Questionnaires and performance tasks may be useful for program evaluation and practice improvement, provided certain challenges with existing measures are resolved through R&D.

This brief was edited by David Yeager, Co-Chair of the Mindset Scholars Network. The author is also grateful for comments provided by Thomas Toch.

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

[ii] Duckworth, A. L., Tsukayama, E., & May, H. (2010). Establishing causality using longitudinal hierarchical linear modeling: An illustration predicting achievement from self-control. *Social Psychological and Personality Science, 1* (4), 311-317. Roberts, B. W., Kuncel, N. R., Shiner, R., Caspi, A., & Goldberg, L. R. (2007). The power of personality: The comparative validity of personality traits, socioeconomic status, and cognitive ability for predicting important life outcomes. *Perspectives on Psychological Science, 2*(4), 313-345.

[iii] The same issue could be observed with teacher-report questionnaires. Tuttle, C. C., Gill, B., Gleason, P., Knechtel, V., Nichols-Barrer, I., & Resch, A. (2013). *KIPP Middle Schools: Impacts on achievement and other outcomes. Final Report.* Mathematica Policy Research, Inc. Dobbie, W., & Fryer Jr., R. G. (2013). *The medium-term impacts of high-achieving charter schools on non-test score outcomes* (No. w19581). National Bureau of Economic Research. Egalite, A. J., Mills, J. N., & Greene, J. P. (2014). *The softer side of learning: measuring students' non-cognitive skills* (No. 2014-03). EDRE Working Paper.

[iv] Performance tasks require carefully trained administrators and can take up to 20 minutes to conduct with a student.

[v] American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1999). *Standards for educational and psychological testing.* Washington, D.C.: American Educational Research Association. See also: Revised Edition, 2014.

[vi] Tuttle et al., 2013. West, M. R., Kraft, M. A., Finn, A. S., Martin, R. E., Duckworth, A. L., Gabrieli, C. F. O., & Gabrieli, J. D. E. (2015). *Promise and paradox: Measuring students' non-cognitive skills and the impact of schooling.* Under review.

[vii] Raudenbush, S. W. (2013). *What do we know about using value-added to compare teachers who work in different schools?* Stanford, CA: Carnegie Knowledge Network. Retrieved from www.carnegieknowledgenetwork.org/knowledge-briefs/.

[viii] Kovacs, M.K. (1992). *Children's Depression Inventory-Short Form* (CDI). New York, NY: Multi-Health Systems Inc.

[ix] Carr, P. B., & Steele, C. M. (2009). Stereotype threat and inflexible perseverance in problem solving. *Journal of Experimental Social Psychology, 45*(4), 853-859. Inzlicht, M., & Kang, S. K. (2010). Stereotype threat spillover: How coping with threats to social identity affects aggression, eating, decision making, and attention. *Journal of Personality and Social Psychology, 99*(3), 467. Inzlicht, M., McKay, L., & Aronson, J. (2006). Stigma as ego depletion: How being the target of prejudice affects self-control. *Psychological Science, 17*(3), 262-269.

[x] Bryk, A. S., Gomez, L. M., Grunow, A., & LeMahieu, P. G. (2015). *Learning to improve: How America's schools can get better at getting better.* Cambridge, MA: Harvard Education Press. 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. San Francisco, CA: Jossey-Bass. Yeager, D. S. & Bryk, A. S. (2014). *Practical measurement.* Unpublished manuscript, Department of Psychology, University of Texas at Austin, Austin, TX.