Evolving Assessments for a 21st Century Education

Maya Bialik
Jonathan Martin
Merrilea Mayo
Bernie Trilling

February 15th, 2016
contact: info@curriculumredesign.org

With gratefulness to our funders:

The Ellen Koshland Family Fund
TABLE OF CONTENTS:

I. EXECUTIVE SUMMARY ........................................................................................................... 2
II. INTRODUCTION ....................................................................................................................... 3
III. WHY WE ASSESS ................................................................................................................... 4
IV. WHAT WE ASSESS ................................................................................................................ 7
V. HOW WE ASSESS ................................................................................................................... 9
   A. CURRENT CHALLENGES ........................................................................................................ 9
   B. SOURCES OF EVIDENCE AND THEIR APPROPRIATE INTERPRETATION ............................... 11
   C. PROMISING AVENUES FOR ASSESSMENT IMPROVEMENT .................................................. 15
VI. FINDINGS IN K-12 AND HIGHER ED .................................................................................. 70
   A. SKILLS .................................................................................................................................. 70
   B. CHARACTER .......................................................................................................................... 76
   C. META-LEARNING .................................................................................................................. 86
VII. FINDINGS IN WORKFORCE READINESS
   A. SKILLS .................................................................................................................................. 70
   B. CHARACTER .......................................................................................................................... 76
   C. META-LEARNING .................................................................................................................. 86
VIII. RECOMMENDATIONS .......................................................................................................... 93
IX. CONCLUSION ......................................................................................................................... 100
X. APPENDICES ........................................................................................................................ 101
I. EXECUTIVE SUMMARY
This preliminary report from the Advisory Panel to CCR’s Assessment Research Consortium (ARC) has three goals:

1. To offer a preliminary set of foundational “Why, What and How” perspectives on educational and workplace assessments for ARC Members to consider, hopefully enabling an easier path toward a clear, common understanding from which to base future ARC activities
2. To provide a preliminary “assessment of assessments” of twelve elements from CCR’s Four-Dimensional education framework, offering detailed analyses of existing higher quality assessment instruments and summary overviews of the current state of educational and workplace assessment for the framework’s Skills, Character qualities and Meta-Learning strategies
3. To offer an initial set of findings and recommendations for consideration by ARC Members, based on the Advisory Panel’s research

The results of this report are a set of nine general recommendations and four specific recommendations for the future of ARC.

General Recommendations:
1. Research Learning Model Competencies and Learning Progressions
2. Further Research the Alignments and Gaps Between Existing Assessment Methods and the Various Purposes of Assessments
3. Adopt/Adapt/Create a Research-based Process for Developing New Assessments
4. Develop a Plan for Dramatically Increasing the Use of High Quality, Authentic Performance Assessments and Comparable Rubrics
5. Create Systems of Assessment Involving Triangulation of Multiple Assessment Methods and Integrated Visualizations of Learning Progress
6. Ensure that All Assessments Provide Guidance for Performance and Practice Improvement
7. Promote the Research and Development of a Wider array of Simulation and Game-based, Technology-empowered “Stealth” Assessments
8. Develop Guidance for Effective 21st Century Assessment Implementation
9. Analyze, Adapt and Prioritize OECD’s and other Policy Suggestions for Meeting the 21st Century Assessment Challenge

Specific Recommendations:
1. Develop an R&D Agenda that Prioritizes the Twelve Elements of the CCR Framework, Identifies which Organizations would be the Best Candidates for the R&D Work, and Includes a Resource Development Strategy for Ongoing Support
2. Create a Living Online Repository of High-Quality Assessments Aligned to the Four-Dimensional Framework.
3. Create a Plan for Technology-enabled Teacher, Trainer and Leadership Development to Align Curriculum, Assessment Implementation, Professional Development, Infrastructure and Policy with Four-Dimensional 21st Century Assessments
4. Develop a Public Education, Outreach and Public Relations Campaign to Inform and Promote Four-Dimensional 21st Century Learning and Assessments
II. INTRODUCTION

The work of the CCR Assessment Research Consortium Advisory Panel

In August 2015 the Center for Curriculum Redesign (CCR), as part of its work toward redesigning the global goals of a 21st century education, launched a preliminary review project in support of its Assessment Research Consortium (ARC) initiative. The goal of this survey project was to “assess the state and future of assessment” in relation to the elements of CCR’s Four-Dimensional Education framework (as detailed in Fadel, Bialik & Trilling, 2015, Four-Dimensional Education: The Competencies Learners Need to Succeed, CCR).

[Also see this report’s Appendix 1 for a detailed prospectus on CCR’s ARC initiative.]

The Advisory Panel consists of:

- Maya Bialik, CCR Education Research Manager
- Dr. Merrilea Mayo, CCR Consultant, Founder & CEO of Mayo Enterprises, a consultancy in the areas of innovation, workforce, technology, and the future of learning
- Jonathan Martin, CCR Consultant, Consultant & Writer on 21st century learning and assessment
- Bernie Trilling, CCR Consultant & ARC Advisory Panel Coordinator, Founder & CEO of 21st Century Learning Advisors, a consultancy on 21st century deeper learning, and education and workforce readiness transformations

Each Advisory Panel member focused on one of the CCR framework dimensions and its elements, as well as more general research questions on K-12 educational and workforce assessments as outlined here:

<table>
<thead>
<tr>
<th>Advisory Members</th>
<th>CCR Framework Dimensions*</th>
<th>Other Assessment Research Questions</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bernie Trilling</td>
<td>SKILLS</td>
<td>Competency relationships across all 4 dimensions</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Description of ARC goals, operations, members, etc.</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Summary of research on the future of assessments</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Structure of Advisory report – Why, What, How</td>
</tr>
<tr>
<td>Jonathan Martin</td>
<td>CHARACTER</td>
<td>Compendium of compendia of assessments</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Varieties of assessment methods</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Guide to analyzing operational criteria</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Theory of action for ARC</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Guide to successful assessment implementations</td>
</tr>
<tr>
<td>Maya Bialik</td>
<td>META-LEARNING</td>
<td>Guide to analyzing impact criteria</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Synthesis of assessment principles/practices</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Synthesis of panel findings &amp; recommendations</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Theory of action for ARC</td>
</tr>
<tr>
<td>Merrilea Mayo</td>
<td>SKILLS, CHARACTER &amp; META-LEARNING for WORKFORCE READINESS</td>
<td>Guide to analyzing technical criteria</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Education and workforce terminology crosswalk</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Impact of CCR elements on workforce assessments</td>
</tr>
</tbody>
</table>

* Note that assessing assessments for the fourth CCR dimension, Knowledge, will be a future challenge for the ARC.

This report outlines summary findings and recommendations based on the preliminary survey work of the Advisory Panel.
III. WHY WE ASSESS
A review of the definition, purposes, model and importance of educational and workforce assessment

A. WHAT IS ASSESSMENT?
Though a healthy variety of nuanced definitions is offered in recent reports on the state and future of assessment (see Appendix 2 for brief synopses of a selection of these reports), the simplest and clearest definition of assessment is from the National Research Council’s (NRC) Committee on the Foundations of Assessment (NRC, 2001, Knowing What Students Know: The Science and Design of Educational Assessment):

Educational Assessment seeks to determine how well students are learning and is an integrated part of the quest for improved education. It provides feedback to students, educators, parents, policymakers, and the public about the effectiveness of educational services.

The 2013 report from the Gordon Commission on the Future of Assessment in Education elaborates on the core approach to this essential educational support function.

Assessment is directed at the generation of inferences concerning developed competencies, the processes by which such competencies are developed, and the potential for their development . . . assessment is best structured as a coordinated system focused on the collection of relevant evidence that can be used to support various inferences about human competencies.

Synthesizing these two definitions, the essential nature of assessment can be summarized in the following list of characteristics.

Educational assessment is:
- A coordinated system
- Collections of demonstrated evidence
- Evidence collections supporting inferences and claims
- Claims about how well students are learning
- Claims about learners’ developed and progressing competencies
- Useful and actionable feedback
- Feedback that guides individual and collective learning
- Feedback that supports integrated educational improvement

The Advisory Panel’s recommended definition of educational assessment that guides the findings and recommendations in this report can then be stated as:

Educational assessment is a coordinated system of collections of demonstrated evidence that supports inferences and claims about student learning and learners’ developed and progressing competencies, providing useful feedback to guide individual and collective learning and to support integrated educational improvement.
B. WHAT ARE THE PURPOSES OF ASSESSMENT?
Based on the Panel’s recommended definition and on research analyses of reasons why we assess learning, the following five universal purposes or goals of assessment are offered.

The purposes of educational assessment are for:
1. **Individual/Group Diagnosis** – to track individual and group learning achievements and progress
2. **Performance/Practice Improvement** – to improve learning and instructional effectiveness by providing useful feedback to students, employees and teachers/trainers
3. **Accountability** – to support system accountability evaluations through feedback on performance of schools, districts, workplaces, etc., in order to guide policy decisions
4. **Program Evaluation** – to provide data on program effectiveness to drive selection and improvement of education and training programs
5. **Research** – to support quantitative and qualitative research on how learners learn across contexts and how to improve learning and assessments

The common thread running through all five of these purposes is the need to apply the results of educational and workplace assessments to *improvement decisions and actions* based on reasoned interpretations of the collection of assessment evidence. Examples of applications include decisions about adapting learning experiences to better meet student needs, improving effectiveness and efficiency of educational and training programs and policies, and deepening understanding in learning and performance assessment science. The overall goal is to deepen the relationship between learning and assessment and move from sorting and compliance more towards continuous improvement of all aspects of the learning experience.

C. THE ASSESSMENT MODEL FOR EFFECTIVE EDUCATION TRANSFORMATION
Based on research by the National Research Council’s Committee on the Foundations of Assessment, with additional contributions from other expert researchers (such as Brown & Duguid, 2000, *The Social Life of Information*, Harvard Business Review), the Advisory Panel suggests a “Circled Triangle” assessment model as a useful guide to the future work of CCR’s Assessment Research Consortium:
The center triangle consists of three components:

1. Learning Model of Competency Development – the effective research-supported learning pathways and progressions toward developing proficiency in a specific competency or set of competencies that underlie collections of assessment evidence and their interpretations

2. Collection of Demonstrated Observable Evidence – the design and execution of research-based performance tasks that effectively demonstrate observable levels of development in a specific competency or set of competencies

3. Interpretation of Evidence to Support Claims – the analysis of the collected evidence from assessments to support claims about the learning of individuals or groups, and the levels of confidence in those claims. This can also include suggestions or prescriptions for further individual or collective learning, program, or research improvement, and commentary on appropriate and inappropriate applications of the assessment results

The outer circle represents the appropriate (and inappropriate) external applications and uses of the assessment findings and data. The link between evidence, interpretation and appropriate action is not guaranteed, and must not be taken for granted. Even with high-quality assessments where all three triangle components are carefully designed and executed and are well-aligned with each other, the assessment results can be applied inappropriately to actions which the assessment was not designed to support, or not be applied at all, due to institutional capacity limitations (e.g., lack of leadership coherence or organizational resources).

Another important aspect of the use of assessment results in the outer circle is the need to promote public awareness and understanding of the essential role assessment plays in educational and workplace performance improvement. These wider public education, outreach, and public relations functions are critical in gaining the support needed from all education stakeholders to help move assessment and systems of education more toward 21st century goals and student success.

D. WHY ASSESSMENT IS IMPORTANT TO EDUCATION AND WORK

Assessment is not an isolated function in the education ecosystem nor in the workplace. What is assessed depends on the education goals desired, the curriculum taught, the learning methods used, and much more. In turn, the results of assessments are crucial to supporting individual and group learning; improving curriculum; personalizing teaching and learning methods; and improving a wide variety of programs and policies throughout education systems and workplaces.

In transforming systems for learner and worker success, it is essential to achieve strong supporting alignments between a 21st century system of assessments and learning goals and standards; curriculum and teaching/training practices; professional teacher, trainer and leader development; safe and supportive environments for learning at school, work and beyond; and the appropriate use of learning tools and technologies.
IV. WHAT WE ASSES
The content of assessments, now and looking forward

A. WHAT WE ASSESS NOW
There is a strong consensus that the majority of currently used large scale assessments are not comprehensive enough to measure what really counts for 21st century student and worker success. Historically, large scale assessments have served a sorting role, that is, their function has been to assess how well students had learned subject matter content, or how much proficiency prospective employees had, in order to predict their future performance. This was done to place students and workers in learning and training programs most appropriate for their proficiency level. Accordingly, statistical measures such as predictiveness and test-retest reliability were a central focus of test design, less so the breadth and depth of the content or understanding being assessed. There continues to be widespread use of a wide variety of locally developed assessments used to support learning and improvements in everyday teaching and training, as well as a number of newly developed assessments designed to better fit students’ and teachers’ needs. As assessments and accountability are taking a more central role in guiding change in education systems around the world, and as education and business goals rapidly transform to meet new challenges, we must now re-think what is being assessed.

B. WHAT WE NEED TO ASSESS
Of the recommendations provided in the 2013 Gordon Commission on the Future of Assessment in Education, and the 2014 Pearson Paper “Preparing for a Renaissance in Assessment”, three of the key findings focus on the content of assessments:

- Assessment must fully represent the competencies that the complex world demands.\(^1\)
- Assessments must accommodate the full range of value outcomes (and not just cognitive/academic achievement narrowly defined and narrowly measured).\(^2\)
- What we choose to assess is what will end up being the focus of instruction.\(^3\)

Central to all of the work of the ARC is the CCR Four-Dimensional framework\(^4\), precisely designed to address the first two principal findings. The framework was developed by analyzing and synthesizing the challenges of the 21st century, and supports a multi-dimensional, holistic view of the learner. In aligning assessments to help students and employees meet the challenges of an increasingly complex world, it is important to note that current Knowledge areas covered in curricular subjects will need to be carefully redesigned in order to include modern disciplines as well as a select set of traditional disciplines, with options for interdisciplinary exploration and themes\(^5\) woven throughout. That said, there are three more learning dimensions that assessments will need to measure: Skills, Character, and Meta-learning, as illustrated here:

---

2 A 2014 Pearson Paper by Hill and Barber “Preparing for a Renaissance in Assessment”
4 Fadel, Bialik & Trilling, 2015, Four-Dimensional Education: The Competencies Learners Need to Succeed, CCR
5 The CCR framework Themes include Digital literacy, Environmental literacy, Global literacy, Information literacy, Design thinking, and Systems thinking.
In addition to gaining deeper understandings of relevant Knowledge, learners need to be ready to apply their knowledge to real-world questions and problems through 21st century 4C Skills: Creativity, Critical Thinking, Communication, and Collaboration. They also need to apply their developing Character qualities (Mindfulness, Curiosity, Courage, Resilience, Ethics, and Leadership) to their own learning and lives. Finally, learners need to engage a fourth learning dimension, which amplifies all the other dimensions: Meta-learning, including reflecting and adapting to progress in the other three dimensions, as well as their beliefs and attitudes about their capacity to learn and manage their learning challenges.

Taken together, the CCR framework synthesizes the demands of the 21st century and the needs of a holistic learner into one integrated framework for guiding the future of assessment and learning progress.

The third key finding points to the larger implications of which content is chosen to be assessed. Assessment can no longer be considered in isolation from learning; it deeply affects classroom instruction and workforce training experiences of all learners. Therefore, the content of assessments must have clear value to all learners, and must be supportive of teachers’ and trainers’ learning goals and methods in their classrooms.

These three findings will serve as guiding principles in the development of this report’s recommendations for the future of assessment.
V. HOW WE ASSESS
The methods used to gather evidence of learning

Because of the simple fact that none of us can peer into the mind of another to see learning happening, in order to assess learning we must first ask learners to produce evidence of their learning. Assessment, at its core, is the gathering of this evidence; each type of evidence and its method of collection has strengths and weaknesses. In designing and using assessments, we must be careful to leverage their strengths, without being misled by their inherent weaknesses.

How we design the conditions for demonstrations of learning (tasks, questions, etc.), how the responses are recorded, how the demonstrated evidence is interpreted and used to support claims about students, teachers, schools, or the workplace, and how those claims are then used in decision making, are all parts of the “how” of assessment.

A. CURRENT CHALLENGES

Multiple Choice Tests
The most commonly used format for large scale assessments is the multiple choice test. This has some large advantages: it is efficient, easy to administer, easy to standardize, and less costly to implement. The current focus on test reliability and validity, as well as on the ease and costs of administration, have made this method of assessment very popular.

Tests are designed so that, according to their model of learning assumptions, if learners have mastered a certain knowledge area or competency, they will successfully answer the posed questions. But this assessment format relies on very indirect evidence of learning.

There are five major problems with the multiple choice tests currently being administered. These assessments:

1. Often fail to include a comprehensive set of competencies of a whole learner in the 21st century (as described in the previous section).
2. Often rely on an implicit model of learning progressions, which is not clearly tied to research.
3. The results of these tests are very often used exclusively for sorting purposes, not supporting learning, so instead of enhancing, they can detract from one’s learning and motivation, and can reinforce a fixed mindset orientation toward learning.
4. Are not more predictive of future outcomes than less intrusive measures such as GPA.6
5. Often are not designed around authentic, real-world tasks and situations, as those are frequently considered too difficult to design, analyze, and standardize; instead these high stakes tests are snapshots of achievement that often do not reflect progress or depth of learning, nor do they guide learners and teachers/trainers toward the next steps to improve learning.

Performance/Practice Improvement

The issues described above are even more problematic when it comes to the use of these assessments for performance and practice improvement. There is strong consensus from the major reports on assessment the Advisory Panel reviewed (see Appendix 2) that practice improvement must be considered an essential goal in the future of assessment.

Pearson’s 2014 report by Hill & Barber, *Preparing for a Renaissance in Assessment*\(^7\) states:

- Assessments should provide meaningful information on learning outcomes (and not be overly reliant on levels that reveal little about what a student can do, or provide information too late or too general to be of value for improving learning, or be reduced to just a mean or cut score percentage).
- Assessments should support students and teachers in making use of ongoing feedback to personalize instruction and improve learning and teaching.

RAND Corporation’s 2013 report sponsored by the Hewlett Foundation, *New Assessments, Better Instruction?*\(^8\) suggests:

- Test content and format should mirror high-quality instruction.
- Score reporting should be optimized to foster instructional improvement.

The Gordon Commission’s recommendations include\(^9\):

- Assessment must be transformed to better support teaching, learning, and human development.
- Assessments must be models worthy of attention and energy of teachers and students.
- New assessment resources and tools are needed that better integrate with classroom teaching and learning, and better represent current thinking on how students learn and on changes in the world at large.

A review of a five-year international study of 38 countries’ perspectives on evaluation and assessment\(^10\) led OECD to a number of policy recommendations, including:

- **Focus on improving classroom practices**
  The point of evaluation and assessment is to improve classroom practice and student learning. With this in mind, all types of evaluation and assessment should have educational value and should have practical benefits for those who participate in them, especially students and teachers.

- **Put students at the centre**
  Because the fundamental purpose of evaluation and assessment is to improve student learning, students should be placed at the centre. They should be fully engaged with their learning and empowered to assess their own progress (which is also a key skill for lifelong learning). It is important, too, to monitor broader learning outcomes, including the development of critical thinking, social competencies,

---

\(^7\) A 2014 Pearson Paper by Hill and Barber “Preparing for a Renaissance in Assessment”


\(^10\) OECD, 2013, *Synergies for Better Learning: An International Perspective on Evaluation and Assessment*
engagement with learning and overall well-being. These are not amenable to easy measurement, which is also true of the wide range of factors that shape student learning outcomes. Thus, performance measures should be broad, not narrow, drawing on both quantitative and qualitative data as well as high-quality analysis.

If assessments are not educational experiences themselves, are not composed of “worthy” authentic learning tasks, and students and teachers are not given supportive and actionable feedback based on the results, the intrusion into students’ learning workflow for the sake of unnecessary sorting (which is less predictive than GPA), and other negative unintended consequences, is simply indefensible.

Attitudes of policymakers, schools, parents, and students reflect this. The title of the 2015 Gallup Poll on the public’s attitudes toward public schools was “Testing Doesn’t Measure Up for Americans,” and it reported that 64% of Americans believe children are subjected to too many standardized tests. Among opinions on appropriate measures of school effectiveness, testing was ranked last, and 41% of those polled supported allowing parents to excuse their child from testing. States across the U.S. have instituted policies exempting their students from standardized tests, providing alternative assessment strategies, and even retroactively granting diplomas to those students who were denied them on the basis of test scores. Testing-optional college admission processes have grown to include 850 higher education institutions with test-optional or test-flexible admissions. Inside higher education, the use of alternative forms of assessments such as rubrics and portfolios has risen dramatically.

And yet, the challenges that multiple choice standardized tests were designed to address are still relevant. 21st century systems of assessments must therefore be designed carefully to accomplish their desired purposes, while maximizing authenticity of tasks, validity of claims, and integration with curriculum, such that students, parents, teachers, schools, and policymakers can use the insights generated to improve learning.

B. SOURCES OF EVIDENCE AND THEIR INTERPRETATIONS

Tests
In order for tests to be correctly interpreted and applied to improving education, they must be considered in their context; in terms of the circled triangle model of assessments, the alignment between the learning model, the evidence collected, and the claims being made based on that evidence, all need to be properly

12 This includes strong majorities from all major demographic groups (Black, Hispanic and White) as well as political affiliations (Republican, Democrat and Independent).
15 Colleges and Universities That Do Not Use SAT/ACT Scores for Admitting Substantial Numbers of Students Into Bachelor Degree Programs. <http://www.fairtest.org/university/optional>
17 Practice Improvement, Program Evaluation, Research, Accountability, and Individual Diagnosis; the purposes of assessment outlined in section III.
interpreted. Historically, the focus has been on validity and reliability of tests and their items and versions in isolation. However, recent work has shown that it is far more accurate to evaluate the proposed score interpretations (claims about test takers, teachers, schools or other units of analysis) and its uses (decisions or actions based on these units of analysis), and not the tests or scores themselves. Michael Kane (of the Educational Testing Service)\textsuperscript{18} provides these eight principles for evaluating validity of tests:

1. **It is the proposed score interpretations and uses that are validated and not the test or the test scores.**
2. The validity of a proposed interpretation or use depends on how well the evidence supports the claims being made.
3. *More-ambitious claims require more support* than less-ambitious claims.
4. More-ambitious claims (e.g., construct interpretations) tend to be *more useful* than less-ambitious claims, but they are also *harder to validate.*
5. **Interpretations and uses can change over time** in response to new needs and new understandings, leading to changes in the evidence needed for validation.
6. The evaluation of score uses requires *an evaluation of the consequences of the proposed uses; negative consequences can render a score use unacceptable.*
7. The rejection of a score use does not necessarily invalidate a prior, underlying score interpretation.
8. The validation of the score interpretation on which a score use is based does not validate the score use.

It is important to be rigorous when interpreting the results of assessments. We must consider all of the relevant pieces – the research-based learning model, the scores on the test, and their interpretations and uses as a logically reasoned argument, and evaluate the strength of that argument in each case. We must be careful to uncover our assumptions about what we are assessing in order to not use results in an inappropriate way.

*Classroom Assignments, Projects & Student Work*

Because the recent assessment reform movement was initially framed as external measures of the summative success of learning in schools, classroom assignments and projects were not often considered proper assessments. As we shift our attention to the importance of supporting learning through assessment, it is clear that classroom assignments and student work have crucial roles to play. This kind of evidence of student learning is the most integrated with classroom experience (since it is classroom experience), and has the potential to be highly authentic.

The challenges of using classroom assignments and projects as a source of evidence of learning is that it is difficult to translate this large amount of information about individual students’ learning in a meaningful way for other purposes (such as accountability, program evaluation, etc.).

\textsuperscript{18} Kane, M. T. 2013. Validating the Interpretations and Uses of Test Scores. *Journal of Educational Measurement*, 50(1), 1–73.  
http://doi.org/10.1111/jedm.12000
Portfolios
Similar to classroom assignments and projects, portfolios are directly related to students’ work in the classroom. They have the added benefit of showing change over time, rather than just a snapshot on a particular day, so they can capture more complex and evolving forms of learning and experiences in the curriculum.\(^{19}\) Finally, portfolios have a built-in self-reflection aspect, which engages the meta-learning dimension, helping students to think about their own progress as they create and review their portfolios. There are a few different kinds of portfolios:\(^{20}\)

- **Showcase portfolios.** Showcase portfolios highlight the best products over a particular time period or course. For example, a showcase portfolio in a composition class may include the best examples of different writing genres, such as an essay, a poem, a short story, a biographical piece, or a literary analysis. In a business class, the showcase portfolio may include a resume, sample business letters, a marketing project, and a collaborative assignment that demonstrates the individual's ability to work in a team. Students are often allowed to choose what they believe to be their best work, highlighting their achievements and skills. Showcase reflections typically focus on the strengths of selected pieces and discuss how each met or exceeded required standards.

- **Process portfolios.** Process portfolios, by contrast, concentrate more on the journey of learning rather than the final destination or end products of the learning process. In the composition class, for example, different stages of the process—an outline, first draft, peer and teacher responses, early revisions, and a final edited draft—may be required. A process reflection may discuss why a particular strategy was used, what was useful or ineffective for the individual in the writing process, and how the student went about making progress in the face of difficulty in meeting requirements. A process reflection typically focuses on many aspects of the learning process, including the following: what approaches work best, which are ineffective, information about oneself as a learner, and strategies or approaches to remember in future assignments.

- **Evaluation portfolios.** Evaluation portfolios may vary substantially in their content. Their basic purpose, however, remains to exhibit a series of evaluations over a course and the learning or accomplishments of the student in regard to previously determined criteria or goals. Essentially, this type of portfolio documents tests, observations, records, or other assessment artifacts required for successful completion of the course. A math evaluation portfolio may include tests, quizzes, and written explanations of how one went about solving a problem or determining which formula to use, whereas a science evaluation portfolio might also include laboratory experiments, science project outcomes with photos or other artifacts, and research reports, as well as tests and quizzes. Unlike the showcase portfolio, evaluation portfolios do not simply include the best work, but rather a selection of predetermined evaluations that may also demonstrate students' difficulties and unsuccessful struggles as well as their better work. Students who reflect on why some work was successful and other work was less so continue their learning as they develop their metacognitive skills.

- **Online or e-portfolios.** Online or e-portfolios may be one of the above portfolio types or a combination of different types, a general requirement being that all information and artifacts are somehow accessible online. A number of colleges require students to maintain a virtual portfolio that may include digital, video, or Web-based products. The portfolio assessment process may be linked to a specific course or an entire program. As with all portfolios, students are able to visually track and show their accomplishments to a wide audience.

---

\(^{19}\) Banta et al 2009 as cited in Using Evidence to Improve Student Learning

Portfolio assessment has little or no fakability problems, but it does pose significant challenges for maintaining high inter-rater reliability rates, though very detailed rubrics and strong training can help. While many detailed high quality rubrics are being developed and researched for content areas and 21st century skills (especially the 4 C’s), there are far fewer tasks or rubrics built for the character and meta-learning dimensions. Evaluating portfolios and recording rubric ratings and feedback also takes time, which can limit their regular and efficient use for formative and improvement purposes; however, advances in technology may help lighten this workload.\textsuperscript{21}

\textit{Rubrics}
Functioning as a framework for interpreting student work, rubrics allow for complex sources of evidence of learning, such as classroom assignments, projects and portfolios, to be assessed in a more consistent and reliable way. They also make expected learning progressions visible, which is helpful for students as clear statements of expectations and scaffolding supports for improvement. Rubrics also help in developing stronger relationships with research as rubrics are often updated to reflect our deepening understanding of learning and teaching. By including students in the design of rubrics, teachers can also incorporate elements of reflective Meta-Learning, as well as self-motivation and personal ownership of students’ learning goals.

\textit{Self-report Surveys}
When a particular type of learning is really difficult to see by observing performance and behavior, self-report surveys are used (often as a complement to other assessment evidence, for research, program evaluation and accountability purposes). Like all large scale assessments, self-report surveys are often available in ready-to-use formats, with validity and reliability built into the design (though of course, the results always need to be re-examined in the larger context of the interpretations and uses of the results).

The drawbacks of self-report are straightforward; if students have incentive to perform well on the assessment (such as if they or their teacher or school is being judged based on the results), it is easy to simply report a higher level of achievement. For instance, self-report measures of curiosity and resilience have been demonstrated effective in some situations, but are not effective when there are stakes for either the test-takers or for the educators administering the tool.\textsuperscript{22}

Self-report assessments are also limited by reference bias: how learners rate themselves very often depends on to whom they are comparing themselves, and by how much they know about the attribute they are rating themselves on. For example, if one is surrounded by highly creative peers, one may well rate oneself poorly in this area, even though one may be more creative than someone else surrounded by less creative peers. After one learns more about what creativity entails, and the many ways in which one can improve and demonstrate creativity, one may also rate oneself lower on this scale, even though, by having learned so much, one probably has become at least a bit more creative (or at least more knowledgeable/skillful about how to be more creative).\textsuperscript{23}

\textsuperscript{23} Duckworth, Yeager. 2015., Egalite, Mills, Green. 2015.
C. PROMISING AVENUES FOR ASSESSMENT IMPROVEMENT

In addition to the directions for growth and improvement of each type of assessment described above, there are several promising recent developments worth further research, which may break the current boundaries between formative and summative, formal and informal, sorting and supporting, and practice improvement, accountability, program evaluation, research, and individual diagnosis.

Performance Measures
The next chapter in the evolution of assessments will necessarily involve a much wider emphasis on performance measures as we move to a Four-Dimensional, 21st century assessment approach. Performance assessments are quite familiar to most of us, as Linda Darling-Hammond writes:\textsuperscript{24}

\begin{quote}
Almost every adult in the United States has experienced at least one performance assessment: the driving test that places new drivers into an automobile with a DMV (Dept. of Motor Vehicles) official for a spin around the block and a demonstration of a set of driving maneuvers, including, in some parts of the country, the dreaded parallel parking technique. Few of us would be comfortable handing out licenses to people who have only passed the multiple-choice written test also required by the DMV. We understand the value of this performance assessment as a real-world test of whether a person can actually handle a car on the road. Not only does the test tell us some important things about potential drivers’ skills, we also know that preparing for the test helps improve those skills as potential drivers practice to get better. . .

The test sets a standard toward which everyone must work. Without it, we'd have little assurance about what people can actually do with what they know about cars and road rules, and little leverage to improve actual driving abilities.
\end{quote}

Performance tasks can be defined as “a structured situation in which stimulus materials and a request for information or action are presented to an individual, who generates a response that can be rated for quality using explicit standards. The standards may apply to the final product or the process of creating it.”\textsuperscript{25}

Performance tasks range from constructed response items on a test, to written essay responses to a prompt, to live demonstrations of skills or presentations of project results rated by rubrics, to sets of rubric-rated student work in a portfolio, to portfolio defense presentations in front of an expert panel to qualify for graduation from secondary school.

A growing number of states and countries have made much progress in incorporating performance assessments into their everyday learning and accountability evaluations, more large scale performance assessment instruments are being used, and new ones are being developed. Notable examples include:\textsuperscript{26}

- Assessment instruments and tests: CWRA+, C-PAS, PISA tasks, NAEP writing, science and arts tasks, PARCC/SBAC items, NOCTI career-technical education assessments, ETS WorkKeys tasks

\textsuperscript{26} Ibid.
- National assessment programs: Finland, Sweden, England, Australia, Singapore, Hong Kong
- International assessment programs: International Baccalaureate (IB) Diploma Program

Progress is also being made in the research, development, implementation and training processes for performance assessments in programs like Quality Performance Assessment (QPA), Stanford’s Center for Assessment Learning and Equity (SCALE), The New York Performance Assessment Consortium, and commercial assessment developers such as ACT, ETS, Pearson, Harcourt, McGraw-Hill, Riverside, Measured Progress and others.

There is also growth in the number and quality of digital assessment tools, online embedded performance assessments, and repositories of performance tasks and rubrics, summarized in the 2014 report, Assessing Deeper Learning: A Survey of Performance Assessment and Mastery Tracking Tools.27

Multiple Measures Methods

Drawing from decades of psychology research and work done at Educational Testing Service in the ‘90s and early 2000s, new methods are being developed to assess “hard-to-measure”, beyond purely cognitive, character competencies using a combination of approaches such as self-report, teacher observation, biodata, situational judgment tests, and forced choice. Assessment instruments like the Mission Skills Assessment (MSA), rated highly by RAND corporation researchers28, or the emerging SSATB and ProExam instruments from current and former ETS affiliated researchers, are examples of this approach. These tests can:

- Be time- and cost-effectively deployed
- Return results speedily
- Compare student and school performance to national and other norms
- Be resistant to faking and coaching (more or less so depending on their particular mix of methods)
- Be designed to limit though not necessarily eliminate reference bias problems.

Some schools are already using these multi-measure instruments for program review and improvement; it is expected new iterations of this approach will be ready for use for student diagnosis, student sorting/selection, and program evaluation by 2016 or 2017.

At present, there is no expectation to use them for accountability, but there is no reason they could not work well, better than the currently used self-reports, in such instances as the “balanced scorecards” for

accountability in the California CORE districts.\textsuperscript{29} Multi-method instruments are not, it should be noted, free, or near-free, as many self-report survey tools are; they are also not embedded into the course of learning but discrete, stand-alone, interruptions to the school-day.

\textit{Simulations and Game-based Assessment Systems}

Simulations and games are already being used to measure important skills and competencies. The 2016 National Education Technology Plan (NETP) reports that:

\begin{quote}
Game-based assessment is designed to leverage parallels between video game design and next-generation learning and assessment. Recent research has focused on promising ways that digital learning can support formative assessment practices—including wraparound features such as annotation tools and dashboards—and ways that games can identify more nuanced conclusions about student learning outcomes. \textsuperscript{30}
\end{quote}

Games such as Zoo U. are immersive, fun and challenging, engaging for learners, and provide instant feedback to students, teachers, and schools about both individual and group performance, compared to norms, on a wide set of social and emotional competencies. Currently being used for individual student diagnosis and school-wide improvement, games and simulations are not easily fakeable and do not suffer from reference bias. The creator of Zoo U., who has received more than 30 federally funded grants, believes the tool can be used for high stakes accountability, though it is not currently. \textsuperscript{31}

Gaming applications too are neither free nor cheap. Because of their capacity for engagement, it perhaps represents less of an interruption than other kinds of tests, and game play provides learning experiences as it is simultaneously assessing the student – a good example of assessment as learning.

\textit{Big Data and Learning Analytics}

Sometimes also referred to as “machine learning,” this represent the greatest challenges and perhaps the greatest opportunities. Though privacy concerns abound,\textsuperscript{32} one way to characterize this approach is that there are enormous volumes of data about students that can be collected and analyzed – every mouse click, every eye movement, every word spoken, every word written. Students’ knowledge and competencies are measured in more traditional ways throughout development, and then computer algorithms crunch millions or billions of the collected student data to determine patterns in those data that are most correlated with students performance.\textsuperscript{33} Once a proficiency model is created for a competency, the collection of student data itself can be used to determine both individual and group proficiencies, in

\textsuperscript{33} Ming, Norma; Baumer, Eric. 2011. Using Text Mining to Characterize Online Discussion Facilitation. Journal of Asynchronous Learning Networks, (15) 2, 71-109
ways that are entirely embedded into online learning experiences, are difficult or impossible to fake, and highly efficient and immediately available.

Another extension of the so-called “big data” or learning analytics enterprise is one which goes beyond keyboard clicks to collect evidence of student emotional affect via eye-tracking, “facial expressions, heart rate, posture, pupil dilation, and more… then analyzed for signs of student engagement, boredom, or confusion.” 34

Beyond the privacy issues, the underlying technologies must advance further to effectively collect, store, and interpret the enormous amount of available student data. It would seem that if this approach can eventually come to fruition, it can be used in any number of ways, including student diagnostics, sorting and selection, program evaluation, improvement, and accountability.

**Systems of Assessment**

No matter how well any individual instrument or methodology serves multiple purposes, it is highly unlikely that we will arrive anytime soon at a place where a single tool or technique can serve the wide array of needs we have for assessment and measurement. As David Conley of the University of Oregon and author of *College Knowledge*, and Linda Darling Hammond, of Stanford University and former US Undersecretary of Education, explain in their important 2013 monograph “Creating Systems of Assessment for Deeper Learning”, 35

> The needed transformations in curriculum, instruction, and assessment will depend on states moving beyond their current testing systems to new systems of assessment that are able to support the development of deeper learning skills, to generate instructionally useful diagnostic information, and to provide insights about a wider range of student capacities that are actionable by students and inform parents, colleges, employers, and policymakers.

Such systems are usually designed to include both the macro and the micro. That is, the high level, high validity and reliability measurements which provide information to policymakers and researchers, and the “much richer school or classroom assessments that offer more detailed information to guide teachers as they develop curriculum and instruction, and students as they revise their work and set learning goals.”

Developing systems of assessment for accountability can:

1. Increase the probability that schools will use good practices on behalf of students
2. Reduce the likelihood that schools will engage in harmful practices
3. Encourage ongoing assessment on the part of schools and educators to identify, diagnose, and change courses of action that are harmful or ineffective

Important to note about these systems is that they represent something more than just collections of various assessments. What is more important and what is difficult to achieve is that they are *aligned*: the

---

problem Conley and Darling Hammond confront and seek to address is that “Current testing regimes in most states typically lack this kind of coherence and synergy.”

In the monograph, the authors showcase various examples of systems that do function with “coherence and synergy,” though it should be noted that their examples, as interesting as they are, only partially extend themselves broadly enough to measure the full range of essential competencies. See for example the Queensland system in this chart from their work:

<table>
<thead>
<tr>
<th>Curriculum guidance</th>
<th>Pre-Secondary Level</th>
<th>Senior Level (Grades 11-12)</th>
</tr>
</thead>
<tbody>
<tr>
<td>External tests</td>
<td>National tests of literacy and numeracy at grades 3, 5, 7, 9—centrally scored</td>
<td>Queensland Core Skills Test, grade 12</td>
</tr>
<tr>
<td>Locally administered performance tasks</td>
<td>Queensland Comparable Assessment Tasks (QCAT): Common performance tasks at grades 4, 6, and 9—locally scored</td>
<td>Course assessments, outlined in syllabus—locally scored/externally moderated</td>
</tr>
<tr>
<td>Locally developed assessments</td>
<td>Local performance assessment systems—locally scored and externally moderated</td>
<td>Graduation portfolios—locally scored/externally moderated</td>
</tr>
</tbody>
</table>

To suggest how an entity might work to systematize their assessments, the following graphic is provided.
Conley and Darling Hammond acknowledge that their examples do not model the full breadth of assessment they are calling for, and accordingly suggest that systems be expanded to include “learning how to learn” skills.  

In order to “develop means for system learning to support continuous improvement at all levels of the system” the report suggests:

- Involving educators in the development and scoring of assessments so that they deeply learn the standards and have opportunities to share practice
- Means for documenting best practices and disseminating knowledge through online platforms sharing studies and highlighting exemplars, school study visits
- Conferences focused on the sharing and development of practice
- Feedback loops to students, educators, and schools about their work (e.g., through exhibitions, educator evaluation systems, and school quality reviews)
- Collaboration opportunities within and across schools and networks

In summary, a successful system of assessments for accountability includes:

- Multiple measures that are complementary and contribute to a comprehensive picture of the quality of learning in classrooms, schools, school systems, and states
- High-quality assessments that encourage and reflect deeper learning and authentic evidence of student readiness to succeed in college and in work
- Profiles of information about students, teachers, schools, and districts that move beyond a single cut score to a richer set of data that can provide indicators of accomplishment and grist for ongoing improvement

---

36 However by locating it in the “Student Profile,” we believe they may be limiting the potential of the use of the assessment of this dimension.
VI. FINDINGS IN K-12 AND HIGHER ED

Results of the Advisory Panel’s scan of existing elementary, secondary, and some tertiary assessments

The Advisory Panel focused mainly on promising assessments in order to evaluate the state of the art in each area of the CCR framework. For a compendium of assessment compendia, see Appendix 3. These more comprehensive reviews can contribute to a much larger repository in the future.

A. SKILLS

Findings from an Initial ARC Review of Assessments of the CCR Skills Competencies

Since the Partnership for 21st Century Skills (P21) began its efforts to promote skills competencies in 2002\(^{37}\), attention and understanding has continued to deepen around the educational value of applying essential skills to knowledge and understanding in real world contexts. The worldwide use of the phrase “21st Century Skills” has grown to the point that in 2015, P21 changed its official name to the “Partnership for 21st Century Learning”, signifying that its mission to raise awareness of the importance of skills in learning was a global success, and refocusing its efforts going forward on the integration and effective transformation of the learning and teaching of knowledge, skills, digital and interdisciplinary literacies, and career and life skills, in as many systems of education as possible\(^{38}\).

The Center for Curriculum Redesign (CCR) has taken a fresh look at the evolving needs for various skills, understandings and mindsets in our rapidly changing times and has come up with a redesigned, newly integrated Four-Dimensional Education framework — Knowledge, Skills, Character and Meta-Learning\(^{39}\). The Skills dimension in this framework contains four elements (often called the “4C” skills, as popularized by P21), with each element having a number of sub-components:

- Creativity – creative/divergent thinking, creative problem solving
- Critical Thinking – reasoning/analysis, systems thinking, making judgments/decisions, critical problem solving
- Communication – reading & writing, speaking & listening, media use
- Collaboration -- teamwork, collaborative problem solving

Though there are a wide variety of well-researched multiple choice and short open response tests for reading and critical thinking, skills assessments, by their nature, are highly performance-dependent, requiring students to demonstrate the skill, or show a product students created by using that skill, to effectively measure their skill level. For example, a written knowledge test of “rules of the road” is not enough to judge a person’s driving ability – a performance test of actually driving a car is absolutely necessary!\(^{40}\)

\(^{39}\) Fadel, C., Bialik, M. & Trilling, B. 2015. Four-Dimensional Education: The Competencies Learners Need to Succeed. CCR.
Skills assessments through performance tasks and evaluative rubrics is receiving new attention\textsuperscript{41}, partly in response to a growing tide of dissatisfaction with, and growing awareness of the limitations of, the recent assessment emphasis on more easily measured content knowledge and basic skills using multiple choice tests.\textsuperscript{42} Now that there is a growing global consensus on the need to learn and assess more complex, multi-faceted skills, character qualities, metacognitive and growth mindset abilities necessary for success in the 21\textsuperscript{st} century, the assessment world is increasingly reform-ready, inclined to innovate, and even predicting an assessment “renaissance”\textsuperscript{43} to meet the education needs of our time.

Technological advances are also driving a significant shift in skill assessment practices\textsuperscript{44}. The stampede of a digital zoo of media forms, from games, simulations, and real-time instant communication connecting to virtually anywhere, to the ease of recording video proof of skill proficiencies, textual fluency, and even one’s emotive data from body sensors, all captured on portable devices or sent to “the cloud”, are all dramatically changing the nature of learning and assessment of ever more complex and subtle skills.

Integrating multiple assessment methods, triangulating multiple learning evidence sources, and building more convincing cases of skill attainment is now possible. Assessments such as the \textit{Mission Skills Assessment}\textsuperscript{45} and the forthcoming ETS \textit{Cognitively Based Assessment of, for, and as Learning} (CBAL)\textsuperscript{46} are early examples of the value of these methods, with even more powerful tools in the wings.

The locus of assessment evidence is rapidly shifting from high-stakes, national or state/provincial standardized, summative proof to local, low-stakes, individualized, formative, classroom and online performance ratings, with a much stronger emphasis on supporting students’ learning progressions than on aggregating often simplified and less informative test data for high-level accountability needs.

CCR’s ARC initiative could not come at a better time to help support and give shape to assessment’s new evolutionary advances – the assessment world is rapidly approaching the threshold of a new era of more powerful, effective, and useful measurement methods that offer a more holistic, supportive, and learner-centered approach to assessments FOR and AS learning. The initial Advisory Panel summary reviews of assessments are outlined below in brief \textit{Overview} sections. An \textit{Examples} table follows with short descriptions of selected assessments and their supporting research for each type of assessment method; then a \textit{Review} section that summarizes the formative and summative assessment opportunities, as well as key gaps in assessment practice; and finally a closing section that outlines potential \textit{Next Steps} for future assessment research, development and effective use.

\textsuperscript{41} One promising example is the Performance Assessment Resource Bank of the US Council of Chief State School Officers (CCSSO) Innovation Learning Network (ILN), created by Stanford’s SCALE and SCOPE groups, which includes Performance Tasks and Portfolio Frameworks, http://www.performanceassessmentresourcebank.org/bin/performance-tasks

\textsuperscript{42} A wide variety of reviews exist; one powerfully critical example is: Nichols, S.L. & Berliner, D.C. 2007. \textit{Collateral Damage: How High Stakes Testing Corrupts America’s Schools}. Harvard Education Press.


Creativity
Creative/Divergent Thinking and Creative Problem Solving

Overview
Creativity, next to Critical Thinking and Communication, is one of the more deeply researched areas in the CCR competency framework, with a broad diversity of assessment tools that are applied in a wide variety of learning contexts — despite a still prevalent misperception that “creativity is too difficult to measure”.47 Tests of aspects of creativity and creative thinking have been in existence for decades — the Torrance Tests of Creative Thinking (TTCT), has been widely used in the US and globally since its introduction in the early 1960s, and continues to be popular and positively reviewed. Rubric-based evaluation of creative work is also a well-researched and well-developed format, as are a number of self-report surveys of creativity48.

The rising demand for innovation in business and in solving complex societal problems has highlighted the need for both individual and group assessments of creativity. Recent research on factors supporting group or organizational “creative cultures”, such as diversity among team members and group practices that build creative confidence49 are stimulating new ways to assess group creative potential. Sophisticated assessment methodologies using live demonstrations and sensor-based biodata, media captures of student work, game and simulation-based data50, and other digital evidence sources of learning, all evaluated by explicit performance rubrics, plus multiple-measure, multiple-format assessments are all becoming increasingly available, and will foster a great deal of innovative creativity assessment research and development in the future.

Examples

<table>
<thead>
<tr>
<th>Methods</th>
<th>Assessments</th>
<th>Description</th>
<th>Supporting Research</th>
</tr>
</thead>
<tbody>
<tr>
<td>Compendium</td>
<td>Various</td>
<td>A list with links to 72 creativity assessment instruments from the Center for Creative Learning (CCL)</td>
<td>CCL. Assessing Creativity Index. (Table of links to 72 assessments)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td><a href="http://www.creativelearning.com/free-resources/assessing-creativity-index">http://www.creativelearning.com/free-resources/assessing-creativity-index</a></td>
</tr>
<tr>
<td>Multiple Choice, Open</td>
<td>Remote Associates Test (RAT)</td>
<td>A general test of creative association consisting of 30-40 seemingly unrelated three word prompts where the goal is to the goal is to</td>
<td>Bowden, E.M. &amp; Jung-Beeman, M. 2003. Normative data for one hundred forty-four compound remote associate problems: Short</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th><strong>Creative Products Evaluations</strong></th>
<th><strong>Consensual Assessment Technique (CAT)</strong></th>
<th>Consensual Assessment is a technique involving the evaluation of creative artifacts, relying on independent subjective judgments of individuals familiar with the domain in which the products were made.</th>
<th>Amabile, T.M. 1982. Social psychology of creativity: A consensual assessment technique. <em>Journal of Personality and Social Psychology</em>, 43, 997-1013. <a href="http://psycnet.apa.org/psycinfo/1983-20083-001">http://psycnet.apa.org/psycinfo/1983-20083-001</a></th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Self-Reports</strong></td>
<td><strong>Creative Behavior Inventory (CBI)</strong></td>
<td>A survey of the presence of common creative activities in a person’s life</td>
<td>See above</td>
</tr>
<tr>
<td><strong>Self-Reports</strong></td>
<td><strong>Biographical Inventory of Creative Behaviors (BICB)</strong></td>
<td>A 34 item scale that assesses everyday creativity across a broad range of domains</td>
<td>See above</td>
</tr>
</tbody>
</table>

Responses, etc. come up with a word that relates to all three insight-like problems with one-word solutions. *Behavioral Research, Methods, Instruments, and Computers*, 35(4), 634-639. http://groups.psych.northwestern.edu/mbeeman/pub s/BRMIC_2003_insight_norms.pdf
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Performance Rubrics</td>
<td>Buck Institute for Education (BIE) Creativity Rubrics</td>
<td>Project based performance rubrics for Grades K-2, 3-5, 6-8, 9-12</td>
<td>BIE. 2015. PBL for 21st Century Success: Teaching Critical Thinking, Collaboration, Communication, and Creativity. BIE. How to Use the 4Cs Rubrics. <a href="http://bie.org/blog/how_to_use_the_4cs_rubrics">http://bie.org/blog/how_to_use_the_4cs_rubrics</a></td>
</tr>
</tbody>
</table>

**Review**

*Formative Opportunities, Assessment for Improvement, and Research & Development*
With a wide variety of assessment instruments and rubric options available to authentically evaluate creative student work, a wealth of selection and constructed response tests for an assortment of specific creative abilities, and a number of well-researched self-report surveys, there are really no excuses for still believing “creativity is too difficult to assess”.

There is a clear need though to develop more group creativity assessments as the “creative culture” of groups and organizations become more and more important to our rising innovation economies. The opportunities for real-time assessments embedded in simulations and games has also opened new channels of evidence gathering that can be further developed using powerful assessment development methods like Evidence Centered Design to bring seamless “stealth assessments” to new generations of digital native students.

**Summative Opportunities and Assessment for Accountability**

Summative assessments of creativity are in very early stages, despite the critical importance of measuring the creative capacity of students, schools, states/provinces, and nations in the 21st century. The Mission Skills Assessment is an early harbinger of more assessments to come that can be used both formatively and summatively at many levels – from student to classroom to school – offering more authentic assessments of creative abilities and targeted support for further developing creative competencies.

**Key Gaps**

There is a clear need to further develop more comparable rubric-based evaluations of creative student work of all kinds – for both creative processes and products – including inventions and innovative solutions to problems (especially important with the growing Maker Movement); and to develop a much wider range of games and simulations with embedded assessments that measure a larger spectrum of creativity abilities in many more domains and interdisciplinary areas of learning.

**Next Steps**

- Promote the use of existing, well-researched creativity assessments
- Use multiple creativity measures to better triangulate evidence of creativity skills development
- Research and develop creativity assessments of groups and teams to better support the need for more innovation in economic development
- Develop and promote the use of comparable rubrics for performance assessments of creative student work, especially creative inventions and physical artifacts measuring both the creative process and product/artifact results
Critical Thinking

Reasoning/Analysis, Judgments/Decisions, Critical Problem Solving, and Systems Thinking

Overview

Critical thinking assessments, like communications assessments, are numerous and varied, and have been a mainstay of both large-scale standardized tests and local formative assessments for a very long time. There are a wide diversity of aspects of critical thinking covered in these tests, such as various forms of reasoning and analysis, decision and judgment making, critical problem solving, and more recently, systems thinking\(^{51}\).

The move toward more complex critical thinking performance tasks and projects with comparable evaluative rubrics is gaining momentum as project based learning approaches expand across the globe. Also more sophisticated assessment methods using digital media captures of demonstrations of critical thinking performances, game and simulation-based data\(^{52}\) evaluated by embedded algorithmic performance rubrics, plus multiple-measure, multiple-format assessments that triangulate diverse learning evidence, will all become increasingly available in the future, and will be rich areas for future critical thinking assessment research and development.

Examples

<table>
<thead>
<tr>
<th>Methods</th>
<th>Assessments</th>
<th>Description</th>
<th>Supporting Research</th>
</tr>
</thead>
<tbody>
<tr>
<td>Multiple Choice, Open Responses, etc.</td>
<td>Watson-Glaser Critical Thinking Appraisal (WGCTA)</td>
<td>One of the most widely used tests of critical thinking in the workplace, but also applicable to educational settings, a multiple choice format test of reasoning from evidence and critical problem solving</td>
<td><a href="http://www.thinkwatson.com/assessments/watson-glaser">http://www.thinkwatson.com/assessments/watson-glaser</a></td>
</tr>
<tr>
<td>Multiple Choice, Open Responses, etc.</td>
<td>Raven’s Progressive Matrices Test</td>
<td>An online test of critical thinking using non-verbal, visual pattern completion challenges where one has to identify the</td>
<td><a href="https://www.raventest.net">https://www.raventest.net</a></td>
</tr>
</tbody>
</table>


<table>
<thead>
<tr>
<th>Category</th>
<th>Test Name</th>
<th>Description</th>
<th>Source</th>
</tr>
</thead>
<tbody>
<tr>
<td>Performance Rubrics</td>
<td>Buck Institute for Education</td>
<td>Project based performance rubrics for Grades K-2, 3-5, 6-8, 9-12</td>
<td>BIE. 2015. <em>PBL for 21st Century Success: Teaching Critical Thinking, Collaboration,</em></td>
</tr>
<tr>
<td>Performance Rubrics</td>
<td>(BIE) Critical Thinking Rubrics</td>
<td>Communication, and Creativity, BIE. How to Use the 4Cs Rubrics. <a href="http://bie.org/blog/how_to_use_the_4cs_rubrics">http://bie.org/blog/how_to_use_the_4cs_rubrics</a></td>
<td></td>
</tr>
<tr>
<td>---------------------</td>
<td>---------------------------------</td>
<td>---------------------------------------------------------------------------------------------------------------</td>
<td></td>
</tr>
<tr>
<td>Multiple Measures</td>
<td>International Baccalaureate (IB) Diploma Programme, Critical Thinking Items</td>
<td>Multiple assessment formats performed internally and externally at IB schools, for 16-19 year olds, covering Critical Thinking, and also Creative Problem Solving, and Communicating skills and six IB-taught subject groups IB Assessments and Exams, <a href="http://www.ibo.org/programmes/diploma-programme/assessment-and-exams/">http://www.ibo.org/programmes/diploma-programme/assessment-and-exams/</a></td>
<td></td>
</tr>
</tbody>
</table>
### Review

**Formative Opportunities, Assessment for Improvement, and Research & Development**

The biggest opportunities for boosting formative support for improving critical thinking skills lie primarily in the rising tide of digital “life stream” data being generated by students’ educational use of smartphones tablets and portable computers. Research and development is needed to capture, analyze and assess this flow of authentic evidence of daily decisions, analyses, conversational reasoning, social judgments, etc., for better insights into how one applies critical thinking skills to everyday life activities.

Better competency, task, and evidence models as part of an Evidence Based Design approach to developing critical thinking skills assessments are sorely needed, as there are so many sub-component skills included within the critical thinking skill category. Developing more widely comparable rubrics for critical thinking performance tasks would also be helpful in unifying formative assessments of critical thinking capabilities.

**Summative Opportunities and Assessment for Accountability**

Along with communication skills, there are no shortages of summative tests and assessments of critical thinking skills, though it is important to make sure that summative assessments also provide rapid feedback for supporting student learning, and not just for sorting accountability purposes –more authentic assessments that can be used both formatively and summatively will be an important area for further development.

**Key Gaps**

More comparable evaluative rubrics for more unified critical thinking performance tasks, more research and development on “stealth assessments” using digital “life stream” data, and more nuanced competency models of the sub-components of critical thinking are all important gaps worth closing.

**Next Steps**

- Identify the key sub-components of critical thinking skills, highlighting which of these sub-skills have the largest positive effects in building critical thinking skills
- Research the relationship between critical and creative thinking skills, and develop a process model that could form the basis of a combined critical and creative thinking assessment
- Use multiple critical thinking skills measures to better triangulate evidence of critical thinking skills improvement
- Create a wider array of simulation- and game-based embedded “stealth” assessments for critical thinking skills
Communication

Reading/Writing, Speaking/Listening, and Media Use

Overview
As two of the three classic “Three R” literacy skills, reading and (w)riting have been at the core of school assessments since schooling began. Appropriate speech and listening skills, though a little less prominent in school assessments than reading and writing, have always been considered essential capabilities for an individual to be considered an “educated person”. As a result of their prominence, there are more and varied assessment instruments for measuring these communication skills, at all levels of education from the individual student on up to international and even the worldwide level (global literacy rate), than any other competencies in the CCR framework.

With the rise of digital communications, new forms of “literacy” have been added to the modern list of essential communication proficiencies – media literacy. High levels of fluency and mastery over the multitude of digital communications options now available – from pictures and video, to animations and virtual worlds, to audio and music libraries — are increasingly indispensable for both digital natives and the dwindling population of digital immigrants (who still speak “technology” with an accent).

And along with the wonders of media technologies, come great opportunities to use those same technologies to modernize and innovate educational assessments. Assessments embedded behind-the-scenes in online learning games, simulations and mobile applications, algorithmic analyses of daily text and audio messages all captured on the digital devices one uses for searching and researching, conversing with classmates, creating artifacts that reflect deepening learning in projects, and both analytical reports and artistic expressions in visual, music or dramatic forms – all of these forms of communication will be accessible evidence of one’s learning pathways and progress.

Examples

<table>
<thead>
<tr>
<th>Methods</th>
<th>Assessments</th>
<th>Description</th>
<th>Supporting Research</th>
</tr>
</thead>
<tbody>
<tr>
<td>Multiple Choice, Open Responses, etc.</td>
<td>National Assessment of Educational Progress (NAEP) Reading &amp; Writing Tests</td>
<td>Since 1986, NAEP assessments have tracked student performance trends by sample testing nationally, by state and in large urban school districts, covering Reading &amp; Writing, and also Math, Science, the Arts, Civics, Economics, Geography, US History and Technology &amp; Engineering</td>
<td>NCES. 2011. Overview of NAEP assessment design. <a href="https://nces.ed.gov/nationsreportcard/tdw/overview/">https://nces.ed.gov/nationsreportcard/tdw/overview/</a></td>
</tr>
<tr>
<td>Performance Tasks/Projects</td>
<td>(In Pilot) Cognitively Based Assessment of, for &amp; as Learning (CBAL) ELA</td>
<td>A multiple methods, performance based assessment based on ECD-developed competency models, with combined formative, summative and teacher development measures of English Language Arts skills of Writing &amp; Reading, including interpretation,</td>
<td>ETS. 2011. CBAL English Language Arts literacy framework. <a href="https://www.ets.org/s/research/pdf/ela_literacy_framework.pdf">https://www.ets.org/s/research/pdf/ela_literacy_framework.pdf</a></td>
</tr>
<tr>
<td>Performance Rubrics</td>
<td>The Conversational Skills Rating Scale (CSRS)</td>
<td></td>
<td>A performance rating form for Conversational Speaking, with evaluations from a the speaker, the partner, and an observer, covering 25 criteria such as speaking rate, fluency, posture, etc.</td>
</tr>
<tr>
<td>Performance Rubrics</td>
<td>Buck Institute for Education</td>
<td></td>
<td>Project based performance rubrics for Grades K-2, 3-5, 6-8, 9-12</td>
</tr>
<tr>
<td>-------------------------</td>
<td>-------------------------------------------------</td>
<td>-------------------------------------</td>
<td>-----------------------------------------------------------------------------------</td>
</tr>
<tr>
<td><strong>Multiple Measures</strong></td>
<td><strong>International Baccalaureate (IB) Diploma Programme</strong></td>
<td>Multiple assessment formats performed internally and externally at IB schools, for 16-19 year olds, covering Communicating, and also Critical Thinking, and Creative Problem Solving skills and six IB-taught subject groups</td>
<td>IB Assessments and Exams, <a href="http://www.ibo.org/programmes/diploma-programme/assessment-and-exams/">http://www.ibo.org/programmes/diploma-programme/assessment-and-exams/</a></td>
</tr>
</tbody>
</table>
Review

**Formative Opportunities, Assessment for Improvement, and Research & Development**

A major opportunity for the formative use of communications assessments is in the area of student media use and the impact of new digital communication methods on student learning. Assessing the relationships between everyday communication uses such as texting, searching, commenting, multi-tasking, etc., and one’s effectiveness in using productive learning strategies to reach one’s educational goals is an important area for further research.

Other areas of research and development for improving the formative assessment of communication skills include the use of more comparable rubrics for performance assessments of communication skills, better competency models of communication skills sub-components (especially in the developmental pathways for reading and writing), the creation of more “stealth” assessments with embedded evaluations of communication skills, and the use of everyday communications evidence captured on digital devices for more authentic analyses of how students apply communication skills to daily life.

**Summative Opportunities and Assessment for Accountability**

Along with critical thinking skills, there are a plethora of summative tests and assessments of communication skills at all levels from classroom to international assessments, though providing more rapid feedback for supporting student learning, and more authentic assessments that can be used both formatively and summatively will be an important area for further development.

**Key Gaps**

---

More comparable evaluative rubrics for more unified communication performance tasks, more research and development on “stealth assessments” using digital “life stream” data, and more nuanced competency models of the sub-components of communication skills are all important gaps worth closing.

**Next Steps**

- Identify the key sub-components of communication skills, highlighting which of these sub-skills have the largest positive effects in building specific communication skills
- Research and create developmental competency models for reading and writing so that assessments can focus more on the high impact sub-skills at different developmental points in a student’s development
- Research the possible negative and positive relationships between the use of various forms of digital media communications, and the effectiveness of specific learning strategies so that assessments can help guide both media uses and productive learning strategies
- Use multiple communication skills measures to better triangulate evidence of communication skills improvement
Collaboration

Teamwork and Collaborative Problem Solving

Overview
Collaboration skills may be the “black sheep” of the skills assessment world – until recently there has been very little attention paid to this competency, with very few assessments worth noting. With the rise in importance of collaborative work skills, especially for the productive development of creative work products by work teams, there has been an accompanying rise in attention paid to assessing levels of productive and creative collaboration.

A big step in the evolution of measuring collaboration skills has been taken with the 2015 implementation of the PISA Collaborative Problem Solving (CPS) computer-based assessment, where students work collaboratively with a virtual “agent” to solve problems. Future research and development in extending the range of simulation and game-based “stealth” assessments, in assessing the dynamic impacts of team diversity on both collaboration and creativity, and in more use of captured “live” (ala GoPro video) interactions among team members for more authentic assessment evidence will bring more depth and breadth to effective and useful collaboration measures.

Examples

<table>
<thead>
<tr>
<th>Methods</th>
<th>Assessments</th>
<th>Description</th>
<th>Supporting Research</th>
</tr>
</thead>
<tbody>
<tr>
<td>Performance Tasks/Projects</td>
<td><strong>ACT WorkKeys Teamwork Assessment Video-based</strong></td>
<td>Video based teamwork scenarios where decision and action choices are presented to the student, and the actions chosen are assessed for promoting appropriate teamwork skills</td>
<td>ACT. 2015. External Review Addresses Validity of ACT's WorkKeys Program for Use in Selection. <a href="http://www.act.org/workkeys/validity.html">http://www.act.org/workkeys/validity.html</a></td>
</tr>
<tr>
<td>Performance Tasks/Projects</td>
<td><strong>Singapore Group Project Portfolio</strong></td>
<td>A group student project on a topic selected by the group, producing a written report, an oral presentation and a group project file of artifacts, with rubric assessments for communication, and also collaboration, scored by teachers</td>
<td>Reviewed in: RAND &amp; Asia Society, 2013, Measuring 21st century competencies: Guidance for educators, <a href="http://asi%D0%B0%D1%81ociety.org/files/gcen-measuring21cskills.pdf">http://asiасociety.org/files/gcen-measuring21cskills.pdf</a></td>
</tr>
<tr>
<td>Performance</td>
<td><strong>Buck Institute</strong></td>
<td>Project based performance rubrics for</td>
<td>BIE. 2015. PBL for 21st Century Success:</td>
</tr>
<tr>
<td>Rubrics</td>
<td>for Education (BIE) Collaboration Rubrics</td>
<td>Grades K-2, 3-5, 6-8, 9-12</td>
<td>Teaching Critical Thinking, Collaboration, Communication, and Creativity. BIE. How to Use the 4Cs Rubrics. <a href="http://bie.org/blog/how_to_use_the_4cs_rubrics">http://bie.org/blog/how_to_use_the_4cs_rubrics</a></td>
</tr>
<tr>
<td>---------------------------------</td>
<td>-------------------------------------------</td>
<td>-----------------------------</td>
<td>----------------------------------------------------------------------------------------------------------------------------------</td>
</tr>
</tbody>
</table>

**Review**

**Formative Opportunities, Assessment for Improvement, and Research & Development**

The opportunities for improving formative assessment of collaboration skills is to have more high-quality assessments developed. Building better competency models of the subcomponents of collaboration, especially related to conflict resolution styles, will only improve the usefulness of collaboration assessments.

**Summative Opportunities and Assessment for Accountability**
Since there are so few summative assessments of collaboration, other than the new PISA Collaborative Problem Solving assessment, it will likely be that simulation and game based collaborative tasks with embedded assessments will be the largest opportunity for building a rich repository of collaboration assessments for a variety of differing contexts and types of team players.

**Key Gaps**
The lack of collaboration assessments is the largest gap, with the need for more nuanced competency models of collaboration skills and more diverse simulation- and game-based assessment experiences, plus the need for more real-world capturing of team interactions to supply an authentic evidence base for assessing real world collaboration, are areas that addressing to support more effective assessments of collaboration.

**Next Steps**
- Research and develop more high-quality assessments of both virtual and real-world collaborations
- Identify the key sub-components of collaboration skills, highlighting which of these sub-skills have the largest positive effects in building overall collaborative skills
- Research and create competency models of collaboration that incorporate research on conflict resolution styles of team members and the levels of diversity of team members on the capacity to collaborate creatively
- Research the effects of various practices in project management on the capacity of teams to productively and creatively collaborate, and develop assessments related to project managing processes to improve collaboration skills
- Use multiple collaboration skills measures to better triangulate evidence of collaboration skills improvement
Multiple Skills Elements
The Four C’s – Creativity, Critical Thinking, Communications, and Collaboration

Overview
Though there are very few assessments designed to measure all four of the 4C skill elements, there is a growing awareness that the future of assessment will involve triangulating multiple measures across multiple competencies to improve reliability and validity, and to take advantage of new digital forms of assessment evidence, as well as providing a more holistic view of students’ growth and development in clusters of competencies.

The use of Evidence Centered Design (ECD) and other comprehensive development methodologies are playing a wider role in assisting the development of these more integrated skills assessments\(^\text{54}\), and research on the best methods and types of evidence to support skill competencies is also increasing\(^\text{55}\).

Examples

<table>
<thead>
<tr>
<th>Methods</th>
<th>Assessments</th>
<th>Description</th>
<th>Supporting Research</th>
</tr>
</thead>
<tbody>
<tr>
<td>Multiple Formats (Multiple Choice &amp; Performance)</td>
<td><strong>OECD Programme for International Student Assessment (PISA) and PISA for Schools</strong></td>
<td>International assessment of Reading, Math, Science and Collaborative Problem Solving, taken by country samples of 15 year olds</td>
<td>PISA FAQ, <a href="http://www.oecd.org/pisa/pisafaq/">http://www.oecd.org/pisa/pisafaq/</a></td>
</tr>
<tr>
<td>Multiple Formats (Multiple Choice &amp; Performance)</td>
<td><strong>International Baccalaureate (IB) Diploma Programme</strong></td>
<td>Multiple assessment formats performed internally and externally at IB schools, for 16-19 year olds, covering Creative Problem Solving, Critical Thinking, and Communicating skills and six IB-taught subject groups</td>
<td>IB Assessments and Exams, <a href="http://www.ibo.org/programmes/diploma-programme/assessment-and-exams/">http://www.ibo.org/programmes/diploma-programme/assessment-and-exams/</a></td>
</tr>
</tbody>
</table>


\(^\text{55}\) Selected references:
- The Forum for Youth Investment. 2014. From soft skills to hard data: Measuring youth program outcomes.
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Performance Rubrics</td>
<td><strong>BIE 4Cs Rubrics for Grades K-2, 3-5, 6-8, 9-12</strong></td>
<td>Project Based Learning oriented rubrics for the 4Cs covering Creativity, Critical Thinking, Communication and Collaboration</td>
<td>BIE. 2015. PBL for 21st Century Success: Teaching Critical Thinking, Collaboration, Communication, and Creativity. BIE. 2015. How to use the 4Cs Rubrics. <a href="http://bie.org/blog/how_to_use_the_4cs_rubrics">http://bie.org/blog/how_to_use_the_4cs_rubrics</a></td>
</tr>
<tr>
<td>Performance Rubrics</td>
<td><strong>AAC&amp;U VALUE Rubrics</strong></td>
<td>College rubrics applicable to secondary schools, covering Creativity (creative/divergent thinking, creative problem solving), Critical Thinking (reasoning, critical problem solving), Communication (reading/writing) and Collaboration (teamwork)</td>
<td>VALUE Rubric Development Project, <a href="http://www.aacu.org/value">http://www.aacu.org/value</a></td>
</tr>
</tbody>
</table>
**Review**

*Formative Opportunities, Assessment for Improvement, and Research & Development*

Though there are an enormous variety of performance-based local instruments and tools for measuring student progress in individual skills, additional research and development is needed to create a more comprehensive and highly integrated strategy for assessing clusters of skills across multiple subject areas. The development of common rubrics, the application of triangulation across multiple measures to increase validity and reliability of evidence-based competency claims (as in the Mission Skills Assessment), and the use of digital device-based assessment applications (EcoMUVE and SimScientists) are all very promising advances that would benefit from deeper learning model development (ECD-based Domain, Student, Task and Evidence models) and assessment user research.

Longer term work is also needed to take advantage of the increasingly “always on” digital connections of students and their “life stream” of data to better support their learning.

*Summative Opportunities and Assessment for Accountability*

Multiple format- and multiple measures-based assessments at the national and international levels (PISA), and at the school, classroom and student levels (PISA for Schools, IB, CWRA+, MSA), are promising developments in the summative use of large- and medium-scale assessments for skills accountability, and even more useful, for both accountability and improvement purposes.

Combining more authentic support for both accountability and improvement at all education levels, where the aggregation of more authentic, individual performance-based assessments, rolled up to the classroom, school, district, state/province, national and international levels, will help assessment reach deeper levels of both effectiveness and efficiency.

**Key Gaps**

The largest gap in assessing the 4Cs skills is the lack of a base set of more comparable performance tasks and projects, using universal rubrics for each of the 4C skills across a wide variety of knowledge domains and interdisciplinary themes. Stanford’s SCALE is currently developing an online repository of performance tasks and rubrics, but much work will be needed to synthesize common taxonomies and rubrics for international use.

Following from this lack of comparable performance tasks and measures is in the need for the embedding of performance tasks and assessments into everyday curriculum activities, both real and virtual, using digital support tools to help make the capture of daily assessment data more seamless. The ability to aggregate performance rated student work into portfolios with meaningful and comparable assessment strategies is also a gap that will need to be filled.

**Next Steps**

- Research and synthesize a consensus list of the core components of each of the 4C skills, especially for critical thinking and communications, so that a standard list of sub-skills can be the basis for a universal assessment framework
- Research and develop online repositories of performance tasks, projects and rubrics, and of high-quality assessments, using an expert consensus set of evaluation criteria
- Develop a strategy for triangulating multiple measures and formats of assessment so that more useful representations of competency levels and more nuanced recommendations on improving skill clusters and individual skills can be made
- Develop a long-term strategy for using student “digital life stream” data, including video demonstrations of skills for learning assessments and support, that is also respectful of student privacy concerns
- Develop a long-term strategy for embedding assessments into both virtual and visceral student curricular learning activities

Skills Next Steps Summary Table

<table>
<thead>
<tr>
<th>Skills Elements</th>
<th>Next Steps</th>
</tr>
</thead>
</table>
| **Creativity**  | - Promote the use of existing, well-researched creativity assessments  
                  - Use multiple creativity measures to better triangulate evidence of creativity skills development  
                  - Research and develop creativity assessments of groups and teams to better support the need for more innovation in economic development  
                  - Develop and promote the use of comparable rubrics for performance assessments of creative student work, especially creative inventions and physical artifacts measuring both the creative process and product/artifact results |
| **Critical Thinking** | - Identify the key sub-components of critical thinking skills, highlighting which of these sub-skills have the largest positive effects in building critical thinking skills  
                          - Research the relationship between critical and creative thinking skills, and develop a process model that could form the basis of a combined critical and creative thinking assessment  
                          - Use multiple critical thinking skills measures to better triangulate evidence of critical thinking skills improvement  
                          - Create a wider array of simulation- and game-based embedded “stealth” assessments for critical thinking skills |
| **Communication** | - Identify the key sub-components of communication skills, highlighting which of these sub-skills have the largest positive effects in building specific communication skills  
                        - Research and create developmental competency models for reading and writing so that assessments can focus more on the high impact sub-skills at different developmental points in a student’s development  
                        - Research the possible negative and positive relationships between the use of various forms of digital media communications, and the effectiveness of specific learning strategies so that assessments can help guide both media uses and productive learning strategies  
                        - Use multiple communication skills measures to better triangulate evidence of communication skills improvement |
| **Collaboration** | - Research and develop more high-quality assessments of both virtual and real-world collaborations  
                        - Identify the key sub-components of collaboration skills, highlighting which of these sub-skills have the largest positive effects in building overall collaborative skills  
                        - Research and create competency models of collaboration that incorporate research on conflict resolution styles of team members and the levels of diversity of team members on the capacity to collaborate creatively  
                        - Research the effects of various practices in project management on the capacity of teams to productively and creatively collaborate, and develop assessments related to project managing |
| **Multiple Skills** | - Research and synthesize a consensus list of the core components of each of the 4C skills, especially for critical thinking and communications, so that a standard list of sub-skills can be the basis for a universal assessment framework  
- Research and develop online repositories of performance tasks, projects and rubrics, and of high-quality assessments, using an expert consensus set of evaluation criteria  
- Develop a strategy for triangulating multiple measures and formats of assessment so that more useful representations of competency levels and more nuanced recommendations on improving skill clusters and individual skills can be made  
- Develop a long-term strategy for using student “digital life stream” data, including video demonstrations of skills for learning assessments and support, that is also respectful of student privacy concerns  
- Develop a long-term strategy for embedding assessments into both virtual and visceral student curricular learning activities |
B. CHARACTER

Findings from an Initial ARC Review of Assessments of the CCR Character Competencies

Assessing the CCR six character competencies has not been the subject of very much attention or development in the education arena to date; however, psychology has been assessing and measuring many of these competencies for research and low-stakes diagnostic and formative purposes for many years, using in most cases self-report assessments.

Most of these competencies are also currently seeing interesting and even exciting assessment innovations coming from “Team Monkey, Team Butterfly, and Team Unicorn”, in the somewhat silly designations used by NPR education and assessment reporter Anya Kamenetz in her 2015 book, The Test.56

- **Team Monkey** represents the organizational and educational psychologists, many currently or formerly associated with an innovation division of ETS, who are building on decades of psychological research to create fake resistant, multi-method measurement tools.
- **Team Butterfly** is the term used for those who are developing performance assessments and portfolio reviews, using in most cases carefully designed and sometimes validated rubrics and high inter-rater reliability, for assessing student work.
- In **Team Unicorn**, educators, psychologists, and software designers are teaming to create gamified assessments, and though Kamenetz only discusses games assessing skills, some companies already have on the market games assessing some of the CCR character competencies.

Although Butterfly is advancing in the cognitive and skills domains, it is only beginning to touch on some of the character areas, such as ethics.

There are common themes running across the six discussions, in some cases, even redundancies. For instance, the VIA character self-report tool57 is comprehensive, esteemed, and addresses many of the CCR character competencies, either explicitly or with attention to one or more related terms. Some of the 24 character strengths in this model are much more strongly correlated to outcomes than others and border on being highly predictive. For example, "temperance" is strongly related to GPA and school behaviors; and "hope" is strongly related to life satisfaction. There is a substantial body of 3rd party research on this tool, spanning multiple countries, demographics, and experimental conditions. It scored quite highly in our ARC evaluation, and it might be a wise avenue for CCR to consider partnering with VIA or designating it as a useful tool for low stakes and formative deployment.

---

56 Kamenetz, Anya. The Test: Why Our Schools are Obsessed with Standardized Testing But You Don't Have to Be. PublicAffairs, 2015.
It should also be noted that, as part of a NCLB waiver accountability index, the non-profit organization Transforming Education\(^5^8\) and the California Office to Reform Education (CORE) are already implementing student self-report measures of self-efficacy, growth mindset, social awareness, and self-management. They are promising to publish research on the survey in the near future.\(^5^9\)

However, the field is not standing still in a reliance upon self-reporting. Consistent across most of the character qualities is that though there exist few tests or highly fake-resistant assessment tools at present, some companies are working hard to develop better such tools. As noted above, (Team Monkey), an innovation center at ETS has developed a multi-trait, multi-method tool called the Mission Skills Assessment for the Independent School Data Exchange (INDEX).\(^6^0\) The MSA provides schools measurements of resilience, ethics, and curiosity, among others. RAND Corporation scientists, including noncognitive measurement expert Brian Stecher, rated MSA as the best such tool available in a late 2013 review entitled Measuring 21st century competencies.\(^6^1\)

That same ETS center is developing a new tool for ARC member SSATB which will use forced choice and other methods to generate individual scored results, in a tool called the Character Skills Assessment\(^6^2\). Meanwhile, scientists from among the ETS group who originally built the MSA are at a new company (ProExam) and are now developing a new multi-trait multi-method tool which will measure several of the CCR competencies.\(^6^3\).

Each of the six competencies is discussed below in an overview section, a Methods section with select assessments, and a Review section that summarizes formative and summative opportunities as well as key gaps, and finally, potential Next Steps for future work.

\(^5^8\) http://www.transformingeducation.org/
\(^6^0\) http://indexgroups.org/msa/
\(^6^2\) See http://www.ssat.org/test/CSA
\(^6^3\) See http://connect.proexam.org/blog/the-demand-is-clear-next-generation-noncognitive-assessment-is-needed-now
Mindfulness

Wisdom, self-awareness, self-management self-actualization, observation, reflection, consciousness, compassion, gratitude, empathy, caring, growth, vision, insight, equanimity, happiness, presence, authenticity, listening, sharing, interconnectedness, interdependence, oneness, acceptance, beauty, sensibility, patience, tranquility, balance, spirituality, existentiality, social awareness, cross-cultural awareness, etc.

Overview

Mindfulness, as characterized in CCR’s framework, spreads more widely across many different areas of interest to educators and assessors than do most other competencies. Not only are there many mindfulness measurements, mostly self-reports and mostly used for adults but potentially adaptable to K-12, but there are also well developed tools for happiness, appreciation of beauty, empathy, social awareness, and cross cultural awareness — each of which is considered in the itemization which follows. Spirituality, and spiritual development, is something which is measured in some religious schools, and could be explored further if it is of interest, but for most secular and public school settings, spiritual development assessment would not be of interest or value. It would seem that one important priority for CCR is to determine which among these many, disparate mindfulness related terms should be prioritized for assessments, and proceed accordingly.

Methods

Self-Report

- Mindfulness Surveys: largely designed for adults; for more details see Workforce Readiness section.
  - Five Facet Mindfulness Questionnaire (FFMQ)
  - Mindful Attention and Awareness Scale (MAAS)
  - Toronto Mindfulness Scale (TMS)
  - The Revised 12-item Cognitive and Affective Mindfulness Scale (CAMS-R)

- Values In Action (VIA)\(^{64}\) — Mindfulness corresponds to appreciation of beauty and gratitude, and grouped together in a category labeled “transcendence.”

---

The VIA character self-assessment is rated well in the literature; its validity and reliability are rated by our evaluation in the middle category of technical quality. It is free and easy to use, and provides helpful resource materials for students and educators.

- **Empathy Surveys**
  Multiple empathy self-assessment surveys exist, with the Toronto Empathy Questionnaire appearing to be best known and most widely used. It is moderately well validated with other surveys and negatively correlated with autism symptomatology.

- **Happiness Surveys**
  Multiple happiness surveys exist, with the best known probably being that of U.Penn’s Marty Seligman. Seligman’s own research has found valid correlations between his survey and other life outcomes, including work productivity.

- **Cross cultural awareness surveys**, some of which also entail short answer essays, biodata, and other ratings.
  - Global Competence Aptitude Assessment
  - Intercultural Effectiveness Scale
  - Global Perspective Inventory
  - Intercultural Development Inventory
  This assessment seems especially on point and has been very thoroughly researched and validated, correlated with success in the workforce and in international study.

*Performance Tasks and Gamification*

- **The MSCEIT YV**, associated with Yale’s Center for Emotional Intelligence and Marc Brackett, Ph.D, measures empathy, broken down into four components:
  - Perceiving Emotions - the ability to identify emotions,
  - Facilitating Thought - the use of emotions to aid and foster ideas,
  - Understanding Emotions - knowledge about emotions, and
  - Managing Emotions - the capacity for emotional regulation.
  It does so not by self-report but by how well people perform tasks and solve emotional problems. It has been thoroughly researched and validated, with correlations to healthier psychological functioning and greater social competence based on both teacher and student ratings, as well as to academic performance in English language arts.

- **Zoo U.**, which measures Empathy and Emotional Regulation.
  This engaging online game, produced by 3C institute in North Carolina by a group of Ph.D. psychologists and game designers, measures in elementary students (and in a companion game, middle schoolers), emotional regulation, impulse control, communication, empathy, cooperation,

---


67 [http://beta.3cisd.com/?portfolio=zoo-u-the-only-research-proven-online-game-that-assesses-and-teaches-social-skills](http://beta.3cisd.com/?portfolio=zoo-u-the-only-research-proven-online-game-that-assesses-and-teaches-social-skills)
and social initiation. It has been researched internally and generated moderately good results for reliability and validity, correlated with school outcomes, GPA, and discipline.  

**Rubrics**

- The Association of American Colleges & Universities (AAC&U) has what they label their VALUE rubrics for college student assessment, including one for Global Learning and for Intercultural Knowledge and Competency. They are still researching interrater reliability and validity of their rubrics. Initial, small studies find good reliability, with validity still being studied.

**Review**

**Formative Opportunities, Assessment for Improvement, and Research & Development**

Mindfulness is among the broadest of the character competencies, and is inclusive of multiple individual areas of research and assessment design. It is clear that there is a lot of activity in this field, and tools are increasingly available to measure mindfulness itself as well as some of its key related competencies such as appreciation of beauty, happiness, empathy, and cross cultural awareness. These robust, often well-researched tools can be, and in some cases already are, used in school settings for formative and low stakes uses.

**Summative Opportunities and Assessment for Accountability**

Although most of the domains within Mindfulness do not appear to be currently used in high stakes or accountability measurements, there are emerging tools and techniques which could be at least potentially deployed this way with moderately good faking safeguards. Rubrics for authentic work, accompanied by rater trainings, are well underway at the college level, and could migrate to K-12. Gamification is happening in Zoo U, and its creator very much believes it is well on its way to being ready for high stakes accountability measurement. Finally the MSCEIT-YV offers another model for how this might work for empathy and emotion perception/understanding.

**Key Gaps**

There is something of a gap for mindfulness per se in the higher stakes accountability realm, but because there does not seem to be demand for or interest in measurement of this particular component of the competency in this regard, one could argue it is not so much a gap per se, as it is not being sought.

Interestingly, and in contrast to most other character competencies, there are quality accountability measurements, and hence relatively less of a gap, emerging for some of the key related terms for mindfulness, including empathy, emotional regulation, and intercultural competency. Indeed, this area can point the way to how performance tasks, games, and rubrics might be valuable resources and methods for better assessing key character qualities in higher stakes arenas.

---

68 Zoo U. An Interactive Adventure for Kids: The only research-proven online game that assesses and teaches social and emotional skills http://personalizedlearninggames.com/wp-content/uploads/2015/07/15-0019-Zoo-U-PLG-WhatsItYv_FINAL.pdf

**Next Steps**

- Review more closely various VIA and other self-report surveys, and establish the ARC recommended measurements for low stakes/improvement research.
- Track and monitor emerging gaming, rubrics, and performance task assessments, looking for new research for reliability and validity, for various related components of mindfulness.
- Look to the existing games, rubrics, and tasks as models for how better fake-resistant assessments might be developed for other character competencies.
Curiosity

Open-mindedness, exploration, passion, self-direction, motivation, initiative, innovation, enthusiasm, wonder, appreciation, spontaneity etc.

Overview

Curiosity in itself does not seem to have been a common area of assessment development historically, though psychology as a field has generated many self-report surveys for the closely related construct of intrinsic motivation. The two terms are treated so closely, that the Mission Skills Assessment\(^7^0\) uses them interchangeably for one of its six constructs.

Methods

Self-reports

- Values In Action (VIA)\(^7^1\) — expressly for curiosity
  The VIA character self-assessment is rated well in the literature; its validity and reliability are rated by our evaluation in the middle category of technical quality. It is free and easy to use, and provides helpful resource materials for students and educators.

- Motivation self-report surveys, particularly intrinsic motivation
  - Choate Self-Assessment
    The Choate Self-Assessment measures intrinsic motivation self-reporting for high stakes selecting (sorting), and has determined in peer reviewed journal published research moderately strong validity for the instrument.\(^7^2\)
  - Motivated Strategies for Learning Questionnaire (MSLQ)\(^7^3\) has been used for many years and has been found to have moderate reliability and validity.
  - Academic Motivation Scale: Similar to the MSLQ, this assessment has been used for many years and has been found to have moderate reliability and validity.
  - Children's Academic Intrinsic Motivation Inventory\(^7^5\)

---


CAIMI items are based on theories of intrinsic motivation measuring enjoyment of learning; an orientation toward mastery; curiosity; persistence; and the learning of challenging, difficult, and novel tasks. It is a self-report instrument consisting of 44 items, to which children rate their agreement or disagreement. There are five subscales, four being subject-area specific (reading, math, social studies, and science) and one addressing school in general. It is rated as having a moderate degree of validity for achievement, IQ, and perception of competence, and inversely related to anxiety and a strong internal consistency.

- Big Five self-report surveys, particularly Openness
  The Big Five Inventory (BFI) is an easy to use self-report survey, and has good validity and reliability ratings.\(^76\) The Openness factor can be seen as a fairly close analogue to curiosity and intrinsic motivation. It is free, and used very commonly in workforce assessments.

**MTMM and Forced Choice methodologies**

- Mission Skills Assessment
  The MSA uses a combination of self-report, teacher rating, and a third section, usually Situational Judgment Tasks (SJT’s), in its evaluation of groups of students. It thereby corrects for, at least in part, the limitations of self-reports. Curiosity is one of the six “skills” assessed in the MSA, which has been rated by the RAND corporation as having moderately good validity and reliability; it is valid for predicting both student GPA and life satisfaction.\(^77\) While easy to administer, it is not very affordable, and is still a new tool in development.

- ProExam and SSATB tools in development for 2016-17
  Both ProExam and SSATB have tools in development; the SSATB “Character Skills Assessment” is expected to measure in individual students some factor of curiosity, using Forced choice methodology and SJTs, and ProExam to measure openness using again, Forced choice and SJT’s. These tools will be carefully designed and developed to be psychometrically sound, though that remains to be proven; they will not be particularly affordable or altogether easy to administer. Their educational impact remains to be seen.

**Review**

*Formative Opportunities, Assessment for Improvement, and Research & Development*

At present, curiosity can be effectively measured in individuals in no/low stakes environments using self-reports which have been established as reasonably sound. Schools could ask students to use VIA or a Big Five openness survey and gain information about student’s individual disposition, and with a bit of effort, aggregate them into group ratings. The MSA is one of the few tools identified which does provide group


assessments of curiosity expressly, and does so for school improvement monitoring--not for individual
diagnosis. Its use of the MTMM methodology is a worthwhile model or template for this kind of work.

**Summative Opportunities and Assessment for Accountability**
The Choate Self-Assessment measures intrinsic motivation self-reporting for high stakes selecting
(sorting), and has determined in peer reviewed journal published research moderately strong validity for
the instrument. The SSATB tool, the Character Skills Assessment, being built by ETS New Constructs
group (led by Patrick Kyllonen) is worth watching: its ambition is to develop a fake-resistant high stakes
individual assessment which could then be modified for use summatively.

**Key Gaps**
There is a major gap currently for assessing curiosity in high stakes environments, though perhaps the
ETS-SSATB emerging tool will represent progress.

**Next Steps**
- Study and select among the various available curiosity and/or intrinsic motivation self-report
  measurements for designation as a go-to CCR resource for no-stakes formative usage, and some
  moderate degree of improvement usage.
- Monitor the ETS-SSATB project and its research findings, and if it is found to be effective,
  consider how it can serve as a model for design of accountability evaluations of curiosity.
- Support research into forced choice and SJT measurements of curiosity for accountability and
  more effective program improvement monitoring, and monitor development of emerging
  ProExam tool (measuring openness as an analogue of curiosity) of these techniques for individual
diagnosis and program improvement.
Courage

Bravery, determination, fortitude, confidence, risk taking, persistence, toughness, zest, optimism, inspiration, energy, vigor, zeal, cheerfulness, humor etc.

Overview
Courage is in itself the subject of very few assessments in K-12, and seems to not be of interest. Psychology has generated a few tools for related constructs, such as optimism. This is an area ARC will need to consider carefully for (1) whether it is worth prioritizing for development, and (2) if so, how to go about developing tools.

Methods
Self-report
- Values In Action (VIA)\(^78\) — expressly for Bravery (and Zest);
  The VIA character self-assessment is rated well in the literature; its validity and reliability are rated by our evaluation in the middle category of technical quality. It is free and easy to use, and provides helpful resource materials for students and educators.

- Optimism Survey (by Martin Seligman)\(^79\)
  - Optimism is identified as being among the various related constructs to courage, though it doesn’t seem to be one of the closest ones.
  - Seligman’s research on optimism has established it as being of great importance to both happiness and success in life.
  - Upon review, little evidence seems available as to the validity or reliability of this tool.

Review
Formative Opportunities, Assessment for Improvement, and Research & Development
According to the Advisory Panel’s research, only the VIA tool is a good fit here, and could be used in schooling contexts with students, though no such practice is known of.

Summative Opportunities and Assessment for Accountability
None at present.

---

Key Gaps
There is a great absence of tools, or development of tools, for assessing courage in both low stakes and high stakes environments. As with many of the other character competencies, VIA stands in as the best current resource available, but unlike others, this is not an area in which other tools are emerging.

Next steps
- Consider whether VIA can be itself a go-to tool for low stakes usage.
- Examine current risk-taking measurements.
- Consider supporting research into into forced choice/SJT/performance task-gamification measurements of courage for accountability and more effective program improvement monitoring.
Resilience

Perseverance, grit, tenacity, resourcefulness, spunk, self-discipline, effort, diligence, commitment, self-control, self-esteem, confidence, stability, adaptability, dealing with ambiguity, flexibility, feedback, etc.

Overview

Resilience, particularly its cognate grit, is probably the most discussed (in educational circles) of all the six character competencies. Psychology has long studied and measured resilience, and offers many tools for educators to consider employing, including the widely distributed Duckworth Grit survey. Duckworth thoroughly distinguishes grit from self-control, treats self-control as the second area of greatest importance to student success, and is developing multiple measures for it.

Methods

Self-report

- Short Grit Scale (GRIT-S)\(^80\) by Duckworth (Grit, self-control)
  Duckworth’s grit survey has famously (TED talk, Tough’s 2012 book *How Children Succeed*) had remarkably good validity for spelling bee championships, West Point perseverance, etc., though we rate it as a moderate strength, which is still good relative to its simplicity. Operationally it is excellent in ease and affordability, and with resources increasingly available; it is already impacting education in multiple ways.

- Values In Action (VIA)\(^81\) — expressly for Perseverance
  The VIA character self-assessment is rated fairly well in the literature; its validity and reliability are rated by our evaluation in the middle category of technical quality. It is free and easy to use, and provides helpful resource materials for students and educators.

- Sedlacek Noncognitive Questionnaire (NCQ)\(^82\)
  For use for college admission, scholarship awards, and for college student placement/counseling, has one of its eight factors being preference for long term goals over immediate gratification.

---


\(^82\) Ting, Siu-Man Raymond, and William E. Sedlacek. *Validity of the Noncognitive Questionnaire-Revised 2 in predicting the academic success of university freshmen*. Counseling Center, University of Maryland, 2000.
has been demonstrated to have moderate reliability and validity. It is freely available and easy to use; many schools are using versions of it, such as the Phillips Exeter for admission and Big Picture Learning Schools for student goal-setting and growth monitoring.

- Big Five self-report surveys — particularly Emotional Stability (opposite Neuroticism) and sometimes Conscientiousness (persistence). The Big Five Inventory (BFI) is an easy to use self-report survey, and has good validity and reliability ratings.\(^{83}\) The Openness factor can be seen as a fairly close analogue to curiosity and intrinsic motivation. It is free, and used very commonly in workforce assessments.

- Bio-data grit activities grid (Duckworth)
  An alternative to traditional Likert Scale self-report is “bio-data” reporting, in which students report actual life experiences (which can be separately verified), as evidence basis for their competencies. So instead of a student choosing five on a likert scale “Very much like me” to an item such as “I persist through many challenges to get things done,” a student might report that she has participated on a swim team with three hours practice daily at 4:30am for six years, and has been injured twice and recovered each time. Duckworth’s website has developed such a tool, a bio-data grit grid. It is a promising approach to strengthening self-reporting, and is being researched more widely for reliability and validity. It is easy to use, uses technology well, and shouldn’t necessarily be very expensive.

**MTMM and Forced Choice methodologies**

- Mission Skills Assessment
  The MSA uses a combination of self-report, teacher rating, and a third section, usually SJT’s, in its evaluation of groups of students; it thereby corrects for, at least in part, the limitations of self-reports. Resilience is one of the six “skills” assessed in the MSA, which has been rated by the RAND corporation as having moderately good validity and reliability; its validity is for both student gpa and life satisfaction. While easy to administer, it is not very affordable, and is still a new tool in development.

- ProExam and SSATB tools in development for 2016-17
  Both ProExam and SSATB have tools in development; the SSATB “Character Skills Assessment” is expected to measure in individual students some factor of resilience or grit, using Forced choice methodology and SJTs, and ProExam to measure emotional stability as a BFF analogue to resilience and grit using self-report, Forced choice, and SJT’s. These tools will be carefully designed and developed to be psychometrically sound, though that remains to be proven; they will not be particularly affordable or altogether easy to administer. Their educational impact remains to be seen.

**Other-Ratings**

- ETS PPI: Personal Potential Index for resilience

Though being discontinued in 2016, this graduate school application (addendum to the GRE) tool developed by ETS and its noncog assessment expert Pat Kyllonen uses a rating form for professors to evaluate students on knowledge and creativity, resilience, communication skills, planning and organization, teamwork, and ethics and integrity. In the policy paper about the PPI tool, Kyllonen reviews the various methodologies to best acquire this information about applicants, and explains why they concluded other-ratings was preferred. The tool uses various formats to strengthen other ratings, including anchoring. It has been found to have moderately strong validity and reliability, and provides a model for potential future tools, though its key drawback continues to be the time required of teachers as assessors.

Performance Tasks

- Duckworth self-control tasks in development
  Duckworth has reported that she is developing self-control performance tasks to measure student competencies; one oft-cited example in her research is where students view a split screen, with math problems on one side, districting visuals/video games on the other side, and then are measured by how frequently they are distracted from their math tasks.

- Zoo U. — expressly for Self Control: Emotional Regulation and Impulse Control
  This engaging online game, produced by 3C institute in North Carolina by a group of Ph.D. psychologists and game designers, measures in elementary students (and in a companion game, middle schoolers), emotional regulation, impulse control, communication, empathy, cooperation, and social initiation. It has been researched internally and generated moderately good results for reliability and validity, correlated with school outcomes, GPA, and discipline.

Review

Formative Opportunities (and/or Measurement for Improvement and Diagnosis)
The good news is that resilience, whether measured by self-report or by MTMM methods, is already being effectively assessed and used for student learning support and improvement efforts. Where stakes are low, and commitment to developing student character is high, schools and districts are using self-report grit surveys to help students strengthen self-awareness, goal-setting, and self-monitoring. The Carnegie Foundation reports in the book Learning to Improve that they are using self-reported “productive persistence” assessments with community college students to measure the effectiveness of their programmatic interventions to strengthen this disposition, and that they have demonstrated validity to their survey.

---

85 http://beta.3cisd.com/?portfolio=zoo-u-the-only-research-proven-online-game-that-assesses-and-teaches-social-skills
In addition to MTMM, Duckworth’s bio-data grid offers another way to strengthen the effectiveness and integrity of self-reporting.

It should be noted that reference bias haunts this field: when students are asked to assess themselves, it becomes complex to determine to whom are they comparing themselves, and when they retake self-assessments, do they actually minimize their own self-ratings having learned more about the construct and its complexity, thereby recognizing more clearly their own limitations. But when used in MTMM, and/or with heavier emphasis on other-ratings, bio-data, and SJTS, and using anchor vignettes to better address reference biases, the current experiments in resilience measurement, in contexts limited in scope and stakes, nevertheless point the way to great opportunities for wider distribution of a strong tool.

**Summative Opportunities (and/or Measurement for Accountability/Program Evaluation)**

At present, there are few summative uses of assessments of resilience. The one exception is that there have been some uses of resilience self-report or other-report surveys for high school, college, and graduate admission selection, though always in very small proportions within a holistic process. That they have shown some moderate success, with supporting research, suggests there is some promise in the opportunity to do more to elevate resilience assessment to accountability measurement.

We should note that Angela Duckworth and David Yeager, probably the two leading academic researchers of resilience and its significance for academic success, warn strongly against using their survey tools for any kind of high stakes assessment.

**Key Gaps**

As with most character competencies, there is a substantial gap around measuring resilience in high stakes environments, though perhaps the ETS-SSATB tool will succeed in this effort and become a model, and the ETS PPI could be another template for the use of other-ratings for this purpose.

**Next Steps**

- Monitor the ETS-SSATB project and its research findings, and if it is found to be effective, consider how it can serve as a model for design of high stakes/accountability evaluations of resilience.
- Monitor development of the emerging ProExam tool usage of these techniques for individual diagnosis and program improvement.
- Support research into forced choice/SJT/performance task-gamification MTMM measurements of resilience for accountability and more effective program improvement monitoring.
Ethics

Benevolence, humaneness, integrity, respect, justice, equity, fairness, kindness, altruism, inclusiveness, tolerance, acceptance, loyalty, honesty, truthfulness, authenticity, genuineness, trustworthiness, decency, consideration, forgiveness, virtue, love, helpfulness, generosity, charity, devotion, belonging, civic-mindedness, citizenship, equality, etc.

Overview

Ethics, though highly important for workforce assessment, has not been much addressed for measurement in the K-12 sector to date. Our initial review finds only the MSA as including a K-12 ethical assessment module. It should be noted that there are tools for measuring a sense of belonging, but they measure student perceptions of their educational environment, and not measurements of any attributes of students as individuals or in groups.

Methods

Self-report

- Values In Action (VIA)\(^{88}\) — expressly for fairness
  
  The VIA character self-assessment is rated fairly well in the literature; its validity and reliability are rated by our evaluation in the middle category of technical quality. It is free and easy to use, and provides helpful resource materials for students and educators.

Other-Ratings

- ETS PPI: Personal Potential Index for ethics and integrity
  
  Though being discontinued in 2016, this graduate school application (addendum to the GRE) tool developed by ETS and its noncog assessment expert Pat Kyllonen uses a rating form for professors to evaluate students on knowledge and creativity, resilience, communication skills, planning and organization, teamwork, and ethics and integrity. In the policy paper about the PPI tool, Kyllonen reviews the various methodologies to best acquire this information about applicants, and explains why they concluded other-ratings was preferred. The tool uses various formats to strengthen other ratings, including anchoring. It has been found to have moderately strong validity and reliability, and provides a model for potential future tools, though its key drawback continues to be the time required of teachers as assessors.

MTMM and Forced Choice methodologies

- Mission Skills Assessment
  The MSA uses a combination of self-report, teacher rating, and a third section, usually SJT’s, in its evaluation of groups of students; it thereby corrects for, at least in part, the limitations of self-reports. Ethics is one of the six “skills” assessed in the MSA, which has been rated by the RAND corporation as having moderately good validity and reliability; its validity is for both student GPA and life satisfaction. While easy to administer, it is not very affordable, and is still a new tool in development.

Rubrics

- AAC&U has what they label their VALUE rubrics for college student assessment, including one for ethical reasoning. They are still researching interrater reliability and validity of their rubrics. Initial, small studies find good reliability, with validity still being studied.

Review

Formative Opportunities, Assessment for Improvement, and Research & Development
As with other areas, there is at least one solid self-assessment easy to use for one aspect of ethics, fairness, although the ethics construct is much broader than fairness alone. Also available are MTMM methods and Rubrics which could be consulted for schools wishing to assess students in ethics and ethical reasoning. As noted, there are corporate sector assessments which perhaps could be adapted for K12 uses.

Summative Opportunities and Assessment for Accountability
There does not appear to be any strong tools available for high stakes or accountability measurements for ethics. The PPI is closest, and serves as a partial model for the practice, but it is designed for college graduates, not K-12 students, is being discontinued, and is onerous to teachers.

Key Gaps
There are big gaps here, in stark contrast to the workforce world; there are only partial tools available even for low stakes measurement, and next to nothing for higher stakes.

Next Steps

- Examine whether workforce ethics assessments are applicable/adaptable to K-12 uses.
- Consider prioritizing development of ethics assessment tools as being of particular strong importance.
- Study whether existing tools (MSA) or emerging tools (ProExam) might be themselves useful or adaptable.
Leadership

Responsibility, abnegation, accountability, dependability, reliability, conscientiousness, selflessness, humbleness, modesty, relationship skills, self-reflection, inspiration, organization, delegation, mentorship, commitment, heroism, charisma, followership, engagement, leading by example, goal-orientation, focus, results orientation, precision, execution, efficiency, negotiation, consistency, socialization, social intelligence, diversity, decorum, etc.

Overview

Leadership as a competency has not historically been a subject of great attention for psychological researchers and educational measurements, but it has recently been addressed in some quality tools from VIA and from SLPI. Accordingly, leadership is now a moderately robust area in the formative arena; the SSATB CSA may potentially add further value in higher stakes applications.

Methods

Self-report

- Values In Action (VIA)\(^9\) — expressly for Leadership. The VIA character self-assessment is rated fairly well in the literature; its validity and reliability are rated by our evaluation in the middle category of technical quality. It is free and easy to use, and provides helpful resource materials for students and educators.

- Sedlacek Noncognitive Questionnaire (NCQ) For use for college admission, scholarship awards, and for college student placement/counseling. One of its eight factors is leadership experience, and has been demonstrated to have moderate reliability and validity. It is freely available and easy to use; many schools are using versions of it, such as the Phillips Exeter for admission and Big Picture Learning Schools for student goal-setting and growth monitoring.

- Big Five Surveys: The Big Five Inventory (BFI) is an easy to use self-report survey, and has good validity and reliability ratings.\(^9\) Leadership in the Big Five is usually seen drawing from facets of

---


Conscientiousness, Extraversion, and Agreeableness, so the BFI can be used to combine those sections and generate a leadership strength profile.

**Combination methodology assessment**
- Student Leadership Practices Inventory (SLPI)
  Combining student self-assessment with other ratings, though there is a self-report only option. This is a quality tool, with strong reliability and validity scoring, founded on quality leadership research. It is easy to use in paper and online, though of course it demands the time of others for rating, but it is not cheap.

- The ETS- SSATB CSA may include a leadership construct.

**Other Ratings**
- ETS PPI: Personal Potential Index for planning and organization
  Though being discontinued in 2016, this graduate school application (addendum to the GRE) tool developed by ETS and its noncog assessment expert Pat Kyllonen uses a rating form for professors to evaluate students on knowledge and creativity, resilience, communication skills, planning and organization, teamwork, and ethics and integrity. In the policy paper about the PPI tool, Kyllonen reviews the various methodologies to best acquire this information about applicants, and explains why they concluded other-ratings was preferred. The tool uses various formats to strengthen other ratings, including anchoring. It has been found to have moderately strong validity and reliability, and provides a model for potential future tools, though its key drawback continues to be the time required of teachers as assessors.

**Review**

**Formative Opportunities, Assessment for Improvement, and Research & Development**
Although we have fewer examples of how leadership assessment is being used in schooling than we do for some of the other character competencies, there do exist some available self-reports, other ratings, and bio-data tools available for schools to use with students for self-evaluation, goal setting, and monitoring.

**Summative Opportunities and Assessment for Accountability**
As with some of the other character competencies, leadership has been used in some very limited higher stakes/accountability ways, for admission selection to high schools, colleges, and graduate schools, and only as a very small proportion of the whole, holistic, decision-making.

**Key Gaps**
There is little work observed being done to develop high stakes leadership assessments, with the possible exception of the ETS-SSATB tool, and with the exception of the discontinued PPI, which could serve as a model.

**Next Steps**
● Study and select among the various available leadership measurements for designation as a go-to CCR resource for no/low-stakes formative usage, and some moderate degree, for improvement usage (the VIA and SLPI are good possibilities here).
● Monitor the ETS-SSATB project and its research findings, and if it is found to be effective, consider how it can serve as a model for design of high stakes/accountability evaluations of leadership.
● Support research into forced choice/SJT/performance task-gamification measurements of leadership for accountability and more effective program improvement monitoring.

Character Next Steps Summary Table

<table>
<thead>
<tr>
<th>Character Elements</th>
<th>Next Steps</th>
</tr>
</thead>
</table>
| Mindfulness        | ● Review more closely various VIA and other self-report surveys, and establish the ARC recommended measurements for low stakes/improvement research.  
● Track and monitor emerging gaming, rubrics, and performance task assessments, looking for new research for reliability and validity, for various related components of mindfulness.  
● Look to the existing games, rubrics, and tasks as models for how better fake-resistant assessments might be developed for other character competencies. |
| Curiosity          | ● Study and select among the various available curiosity and/or intrinsic motivation self-report measurements for designation as a go-to CCR resource for no-stakes formative usage, and some moderate degree of improvement usage.  
● Monitor the ETS-SSATB project and its research findings, and if it is found to be effective, consider how it can serve as a model for design of accountability evaluations of curiosity.  
● Support research into forced choice and SJT measurements of curiosity for accountability and more effective program improvement monitoring, and monitor development of emerging ProExam tool (measuring openness as an analogue of curiosity) of these techniques for individual diagnosis and program improvement. |
| Courage            | ● Consider whether VIA can be itself a go-to tool for low stakes usage.  
● Examine current risk-taking measurements.  
● Consider supporting research into into forced choice/SJT/performance task-gamification measurements of courage for accountability and more effective program improvement monitoring. |
| Resilience         | ● Monitor the ETS-SSATB project and its research findings, and if it is found to be effective, consider how it can serve as a model for design of high stakes/accountability evaluations of resilience.  
● Monitor development of the emerging ProExam tool usage of these techniques for individual diagnosis and program improvement.  
● Support research into forced choice/SJT/performance task-gamification MTMM measurements of resilience for accountability and more effective program improvement monitoring. |
| Ethics             | ● Examine whether workforce ethics assessments are applicable/adaptable to K-12 uses.  
● Consider prioritizing development of ethics assessment tools as being of particular strong importance.  
● Study whether existing tools (MSA) or emerging tools (ProExam) might be themselves useful or
### Leadership

- Study and select among the various available leadership measurements for designation as a go-to CCR resource for no/low-stakes formative usage, and some moderate degree, for improvement usage (the VIA and SLPI are good possibilities here).
- Monitor the ETS-SSATB project and its research findings, and if it is found to be effective, consider how it can serve as a model for design of high stakes/accountability evaluations of leadership.
- Support research into forced choice/SJT/performance task-gamification measurements of leadership for accountability and more effective program improvement monitoring.
C. META-LEARNING

*Findings from an Initial ARC Review of Assessments of the CCR Meta-Learning Competencies*

The fourth dimension of the Four-Dimensional framework lies “meta” to the other three. It concerns the way students reflect on their learning and adapt in order to improve. In this way it can help enhance the other three dimensions and transfer of learning. The two elements outlined here are Metacognition and Growth Mindset.

**Metacognition**

**Methods**

**Self Report**

Metacognition is often split up into metacognitive knowledge and metacognitive skills (prediction, planning, monitoring, and evaluating). The majority of assessments of metacognition come from psychology research, and take the form of self report exercises.

Because metacognition can be both broadly applicable (such as reflecting on problem solving strategies, effort, and motivation) and specifically scaffolding the learning progression of a particular subject, (such as reflection on reading strategies or math strategies), it is important to note that there is a collection of assessments that are being developed specifically for various subjects. Some of the big focuses have been Mathematics, and English as a Foreign Language (including Reading), and Chemistry.

Typically, students indicate on a Likert-type of scale to what extent a statement (e.g., ‘I ask myself questions to make sure I know the material I have been studying’) is representative of their behaviour either before (prospective), or after (retrospective) the main exercises or assessment. These are called “off-line” assessments because they do not take place at the same time as the main task (such as a math test).

Two widely used and reasonably valid, and reliable assessments are:

- Motivated Strategies for Learning Questionnaire (MSLQ) 
  - For ages 11 and up; takes 20-30 minutes to complete; free.
  - Used in “literally hundreds” of educational research projects

---

93 e.g. Abhakorn, M. L. "Investigating the Use of Student Portfolios to Develop Students' Metacognition in English as a Foreign Language Learning." *Journal of Language Teaching and Research* 5.1 (2014): 46-55.
- Contains subscales that can be used separately. This includes measurements of metacognitive self-regulation, planning, rehearsal, and help-seeking.
- Some items begin with a statement such as “Compared with other students in this class.” which may induce reference bias.

- Learning and Study Strategies Inventory (LASSI)\(^97\)
  - For ages 13 and up; takes 15-20 minutes to complete; costs $4 per test.
  - Relevant Subscales: Selecting Main Ideas; Self-Testing (e.g. "I stop periodically while reading and mentally go over or review what was said"); Time Management; and Attitude (e.g. "I feel confused and undecided as to what my educational goals should be")

Both tools report predictive validity and reliability measures that tend to score medium on the Advisory Panel’s proposed Technical Analysis Criteria (see Appendix 4).

There are also “on-line” assessments, such as think aloud protocols, which take place during the task. These are both more accurate and far more time consuming than their offline counterparts.\(^98\) There are some interesting validity issues that come into play when online and off-line assessments are compared.\(^99\)

**Multi-Method**
The trend of using multiple measures to assess metacognition is gaining momentum. This can include evidence from surveys such as those described above, as well as combined surveys such as the The Evaluation and Prediction Assessment (EPA2000),\(^100\) and teacher questionnaires.\(^101\) However, this again takes extra effort on the part of the educator. There are some promising avenues to explore in the use of technology to support and scaffold metacognitive strategies.\(^102\)

**Classroom Work**
Metacognition is rising in importance as educators are realizing its importance and developing learning experiences to complement their existing assignments by encouraging students’ metacognition. “Exam Wrappers,”\(^103\) developed by Marsha C Lovett, ask students to reflect on the questions they got right and wrong, what was effective in their studying and what wasn’t, and prompts them to use this reflection to improve their habits for the next exam.\(^104\) In theory, there is no reason that these should be limited to exams, but can be used to help students reflect on homework, projects, collaborative work, etc. by highlighting the meaningful questions and scaffolding self-reflection. These exercises can be interwoven

---


\(^103\) http://www.cmu.edu/teaching/design/teach/examwrappers/
with the regular classroom activities as needed. For example, there is evidence that writing about one’s worries before a stressful exam boosts performance on the exam.¹⁰⁴

Portfolios
By design, portfolios are meant to encourage and improve metacognition; students must think about their own learning in order to organize and present it, and reflect on their performance and development over time. As classroom assignments, portfolios have a long history as educational tools, however they are difficult to score as assessments. With the rise of ePortfolios, there has been a rising interest in assessing metacognition, and some research has been shown that data can be used as evidence of metacognition.¹⁰⁵

All types of portfolios (showcase, process, and evaluation - described above) can be used to enhance and assess metacognition. They are a very authentic method of assessment, and allow the opportunity for teachers to model metacognitive thinking (e.g. “I think this essay is a great example of how you carefully organized your ideas.”) and to prompt students to reflect on their learning (e.g. “what was effective/non-effective about the learning strategy I chose to use?”). Portfolios can also be used to lead a meeting in which student and teacher co-construct the discussion about the student’s learning, or one in which the students lead the meeting by discussing the work in their portfolio, why they chose it, and what it shows about their learning.

Next Steps
● Support research into learning progressions of metacognition (both general and specific, and model their intersection and interaction). Attend to the discrepancy in on-line and off-line assessments, triangulating the useful assessment methods and conceptual constructs.
● Iteratively develop rubrics that reflect these findings so that educators can use research-based tools of assessment in their classrooms to complement authentic tasks like portfolios and classroom work, and the data can be used to continue improving the rubrics.
● Catalogue the ways in which metacognition can be interwoven into regular classroom activities.
● The generalizability also provides an opportunity for educators to “gather, share, and promptly benefit from data about students’ learning by coordinating their efforts across disciplines, institutions, and countries.”¹⁰⁶
● Explore technology’s potential to enhance metacognition learning and assessment.

Growth Mindset

The work of Carol Dweck describes a Growth Mindset as one in which learners believe that their intelligence (or proficiency potential in a particular subject or competency) is malleable, and with work they will improve. In contrast, a fixed mindset is one in which students believe that intelligence (or proficiency potential in a particular subject or competency) is innate, or “fixed,” and while they can somewhat improve it, ultimately they will never truly master it. This is closely tied with a Mastery orientation (in which learners are motivated by a desire to understand, and thus approach difficulties by applying more effort) and a Performance orientation (in which learners are motivated by demonstrating mastery, and thus approach difficulties by decreasing effort).

While this work has been groundbreaking, it often fails to have the expected effects when applied in schools. Part of this may be that it is oversimplified to mean “just apply more effort” or as a catch-all to explain students’ failure. This points to a breakdown in the circle element of the circled triangle model of assessment.

The reasons are varied. Due to the constraints of time, curriculum, and logistics, students are often not given the chance to struggle through challenges, which is essential for learning and the process in which having a growth mindset and mastery goal orientation is most helpful. Also, assessments that do not include a focus on practice improvement (serving the purpose of sorting students, but not supporting their learning) further engrain a fixed mindset. Finally, teachers’ beliefs about fixed mindsets can influence students’ beliefs. For these reasons, the current effect size of Growth Mindset is .19, much lower than .4, which is considered the “hinge point” (the point at which a year’s input will yield a year’s worth of growth).

Methods

Self-Report

Carol Dweck outlines a simple assessment of Growth Mindset called Implicit Theories of Intelligence Scale. Since then, others have been developed based on the same foundation. The Academic Mindset Meter Assessment, developed at Stanford, is a research-supported survey that measures the mindsets of a group of students. In addition to Growth Mindset, it includes a measure of Belonging, Self-Efficacy, and Relevance/Purpose. This can provide teachers with information about their students’ mindsets and the implications, however, it cannot be administered more than once to a given student.

A key issue with the academic mindset survey is the multitude of ways it is being used, many of which the survey was not designed to handle. In particular, it is being used for classroom-sized studies, even though its statistical properties are not robust enough to be predictive or even informative at such a small scale. The Motivated
Strategies for Learning Questionnaire (MSLQ)\textsuperscript{110} described in the previous section also contains valid, reliable, and widely used subscales related to growth mindset.

As part of new efforts for accountability, the non-profit Transforming Education and the California Office to Reform Education (CORE) districts will be experimenting with new measures of Socio-Emotional Competencies, including Growth Mindset, Self Efficacy (a very closely related concept), and Self-Regulation (closely related to Metacognition)\textsuperscript{111} based off Dweck’s work.\textsuperscript{112}

**Next Steps**
- Monitor and support progress of CORE schools’ assessment techniques and findings
- Investigate measures of teachers’ and administrators’ mindsets, as these are passed to students through consistent signaling much more powerfully than explicit instruction.

**Meta-Learning Next Steps Summary Table**

<table>
<thead>
<tr>
<th>Meta-Learning Elements</th>
<th>Next Steps</th>
</tr>
</thead>
</table>
| Metacognition | ● Support research into learning progressions of metacognition (both general and specific, and model their intersection and interaction). Attend to the discrepancy in online and off-line assessments, triangulating the useful assessment methods and conceptual constructs.  
● Iteratively develop rubrics that reflect these findings so that educators can use research-based tools of assessment in their classrooms to complement authentic tasks like portfolios and classroom work, and the data can be used to continue improving the rubrics.  
● Catalogue the ways in which metacognition can be interwoven into regular classroom activities.  
● The generalizability also provides an opportunity for educators to “gather, share, and promptly benefit from data about students’ learning by coordinating their efforts across disciplines, institutions, and countries.”\textsuperscript{113}  
● Explore technology’s potential to enhance metacognition learning and assessment. |
| Growth Mindset | ● Monitor and support progress of CORE schools’ assessment techniques and findings  
● Investigate measures of teachers’ and administrators’ mindsets, as these are passed to students through consistent signaling much more powerfully than explicit instruction. |

VII. FINDINGS IN WORKFORCE READINESS

CCR and ARC recognize that the preparation and matching of students to jobs after graduation as an important function of educational systems. Methods of selection used in the employment sector were reviewed across the 12 competencies of the CCR framework. Below we provide a summary of the state-of-the-art assessments, and suggestions for future work for each competency (a more in-depth paper will be published separately). At the end of this section there is a summary table for all three dimensions discussed.

The Skills Dimension

Below is a summary table of the Advisory Panel’s findings on the current state of assessment of each skill in the Four-Dimensional framework in the workforce.

<table>
<thead>
<tr>
<th>CCR Element</th>
<th>Importance to Employers</th>
<th>Most Commonly used assessment approaches in Workforce</th>
<th>Explanation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Creativity</td>
<td>Low to Negative (but high for corporate)</td>
<td>Game, Make-a-List Idea Fluency Exercises, Big 5 personality assessment (Likert scale/multiple choice self-report)</td>
<td>Ranked dead last by employers on a major survey on what's important to them in hiring. Also, given that pre-hire personality tests screen for conscientiousness, they will automatically screen against creativity, since these two personality traits are anticorrelated in most of the population. On the other hand, it has been rated as the #1 factor necessary for future success in a global CEO study. The discrepancy might be explained by different characteristics being desired for different levels in the hierarchy.</td>
</tr>
<tr>
<td>Critical Thinking</td>
<td>High</td>
<td>Situational judgment tests such as the Watson-Glaser. However, this test – and others like it – is rarely used at the present time.</td>
<td>Ranked in the top 3 of requested skills by major employer surveys. Disproportionately little testing of this construct in the US due to legal challenges arising from large score differentials between majority and minority ethnic/racial groups. College degree (also racially discriminating) is used as a proxy.</td>
</tr>
<tr>
<td>Communication</td>
<td>High</td>
<td>Degree (and cover letter, writing sample, etc.) is most common. Reading, Writing, Listening and Speaking tests</td>
<td>Also a &quot;top 3&quot; skill in employer surveys, communication is predominantly assessed by proxy - assuming that a college or high school degree equates to an acceptable level of reading, writing, and speaking skill. Secondarily, these skills are assessed via informal in-person means: review of resume and cover letter (to judge writing skill), job interview conversation (gives some indication of oral communication skill), and response to written instructions during the application process (reading skills). A few niche markets, e.g., employees needing to conduct international business, take part in explicit and rigorous communications testing.</td>
</tr>
<tr>
<td>Collaboration</td>
<td>High</td>
<td>Stevens-Campion Teamwork KSA test and/or individual personality traits the employer feels probably relate to teamwork, from Big 5 personality assessments</td>
<td>The last of the &quot;top 3&quot; in-demand skills from employer surveys, this skill has a bewildering array of conceptual models available, but little in the way of workplace-validated assessments.</td>
</tr>
</tbody>
</table>
The following sections provide a brief overview of each of these competencies, and suggest directions for future work. For more depth, see the extended and heavily referenced report *Workforce Assessments of 21st Century Competencies*.

**Creativity**

*Summary of Interest and the State of the Art*

Creativity is not a highly sought after dimension to test. In Big-5 based assessments meant for employee selection, creativity is typically screened against\(^\text{114}\), as it is inversely correlated with conscientiousness. In a few places, creativity appears incidentally as a small factor in tests primarily designed to assess other behaviors or skills. In contrast, at the CEO level, creativity has been claimed\(^\text{115}\) as the most crucial factor for future corporate success. Thus, creativity may have more value for those in charge compared to the average employee, for whom the ability to obediently follow instructions is a higher corporate priority.

<table>
<thead>
<tr>
<th>Test Name</th>
<th>Form Factor</th>
<th>Overarching Workplace Purpose</th>
</tr>
</thead>
<tbody>
<tr>
<td>Virtual Job Tryout</td>
<td>Make-a-List</td>
<td>Pre-hire screening (creativity is screened for in a few isolated occupations)</td>
</tr>
<tr>
<td>Hogan Select</td>
<td>Multiple Choice Self-Report</td>
<td>Pre-hire screening (creativity is screened against)</td>
</tr>
<tr>
<td>Wasabi Waiter (with a companion game, now discontinued)</td>
<td>Gameplay</td>
<td>To predict success in an internal funding competition designed to support innovation</td>
</tr>
<tr>
<td>Founder Institute Assessment</td>
<td>Unknown</td>
<td>To identify future successful entrepreneurs</td>
</tr>
</tbody>
</table>

*Next Steps*

While creativity is not a well-liked attribute for most standardized corporate environments, it appears to be an element in predicting which individuals survive and thrive as entrepreneurs and innovators. Expanding the ability to assess and predict entrepreneurship/innovation would be a major achievement, if deployed on a public platform, rather than held captive on current privately-held platforms. Many countries’ governments place a premium on new business creation within their borders and would benefit from an assessment that indicates the readiness of their populations to innovate. In the US, nearly 100% of net new job creation, and 20% of total job creation, can be attributed to companies less than a year old.\(^\text{116}\)

**Critical Thinking**

*Summary of Interest and the State of the Art*

While employers almost unanimously agree critical thinking is crucial to employee performance, and while the literature backs up this assertion with high statistical correlations between cognitive skill and job performance,\(^\text{117} \ 118 \ 119 \ 120\) testing for critical thinking is not legally straightforward. Because of the  

---

\(^\text{114}\) Insert reference to full-fledged workforce report here


large minority-majority achievement gaps on critical thinking/cognitive ability tests, most U.S. employers are not legally allowed to test critical thinking as a generic skill, and the few that can assess it as job-related skill, must do so using a form factor that replicates the form factor of the job itself. Since few jobs involve taking multiple choice tests, there is a challenge in coming up with a critical thinking exercise that is mass-producible and mass-gradable, yet mimics the daily job duties to the point where it easily passes regulatory compliance. At least one test, Virtual Job Tryouts, has attempted to bridge this gap by including contextual video and dedicated, branded, interactive, company-specific exercises in its job content test offerings. In other words, it tries to replicate the workplace environment as faithfully as possible. However, the expense of this approach has limited it to large retail employers.

<table>
<thead>
<tr>
<th>Test Name</th>
<th>Form Factor</th>
<th>Overarching Workplace Purpose</th>
</tr>
</thead>
<tbody>
<tr>
<td>Watson-Glaser Critical Thinking Test</td>
<td>Situational judgment (multiple choice)</td>
<td>Research on relation of critical thinking to job performance. Also, managerial development.</td>
</tr>
<tr>
<td>Job content tests (e.g., programming test, accounting test)</td>
<td>Computer or paper and pencil</td>
<td>Pre-hire screening</td>
</tr>
<tr>
<td>Virtual Job Tryout</td>
<td>Multimedia, interactive job content test (+ integrated personality test)</td>
<td>Pre-hire screening</td>
</tr>
</tbody>
</table>

**Next Steps**

A number of papers have lamented the grievous “diversity-validity dilemma” described above. One future project ARC may wish to undertake is dissecting the underlying mechanisms of this discrepancy and the necessary changes in the learning model in order to create a critical thinking test that does not result in racially biased score outcomes. A promising line of argument from the literature suggests that the primary obstacle for minorities with respect to any kind of mental ability test is the vocabulary and sentence structure these tests use to pose their problems – not the decision-making process that comes afterwards, which is intended to be the focus of the test. The literature is rife with examples showing critical thinking test scores have extraordinarily high correlations with reading test scores. The level of correlation (r 0.7 to r 0.8) is what one would expect if the same individual were simply taking the exact same test, on the exact same topic, twice in a row.

For intelligence tests, which are close relatives of critical thinking tests, it has been shown that if all students are given a dictionary during the test, the difference in scores between races disappears.

---


Alternatively, replacing words typically known only by Whites, with nonsense words, and then teaching both groups the new words, also completely erases the “intelligence gap”. Modern-day intelligence tests, such as the Siena Reasoning Test, have substituted nonsense words for sophisticated English words, otherwise keeping the problems the same. The result is an “intelligence” test that predicts work performance as accurately as prior tests but has no score differential between black and white students (although still some remaining differential between black and white working adults).

From this historical antecedent, it stands to reason that if one were to similarly erase the vocabulary-dependency of critical thinking tests, one might arrive at a critical thinking test that was both valid and unbiased. The emerging crop of “problem-solving” tests, where intellectual challenges are posed as exercises with diagrams, shapes, game exercises, or computer menus – rather than narrative passages – may be a first step in this direction.

Communication

Summary of Interest and the State of the Art
For the most part, employers use ancillary information from the hiring process to get a sense of applicants’ communication skills: a college/high school degree connotes overall literacy, a cover letter and resume shed light on written communication specifically, and an in-person interview gives a sense of oral communication skills.

Communications assessments that explicitly measure reading, writing, speaking, and listening are relegated primarily to two well-established markets: entry level workers in government workforce training programs and international professionals who need to prove they have a second language proficiency adequate to their intended work overseas. The scale of these markets is large, however. The TOEIC test, for example, is taken by over 5 million people annually; the CPE is taken by about 4 million people annually. The existing tests align their vocabulary and content somewhat to occupational domains as shown below. The assessments are largely computer-delivered and computer-graded, with the exception of speaking, which typically demands a human evaluator. The professional-level writing tests are also hand, rather than computer, graded.

<table>
<thead>
<tr>
<th>Test Name</th>
<th>Primary Market</th>
<th>Overarching Workplace Purpose</th>
</tr>
</thead>
<tbody>
<tr>
<td>Comprehensive Adult Student Assessment System (CASAS)</td>
<td>Government workforce training (assistance) programs</td>
<td>Ensure the individual’s communication skills are adequate for an entry level job.</td>
</tr>
<tr>
<td>Test of Adult Basic Education (TABE)</td>
<td>Government workforce training (assistance) programs</td>
<td>Ensure the individual’s communication skills are adequate for an entry level job.</td>
</tr>
<tr>
<td>Massachusetts Adult Proficiency Test (MAPT)</td>
<td>Government workforce training (assistance) programs</td>
<td>Ensure the individual’s communication skills are adequate for an entry level job.</td>
</tr>
</tbody>
</table>

126 ibid.
128 ibid.
Basic English Skills Test (BEST) | Government workforce training (assistance) programs | Ensure the individual’s communication skills are adequate for an entry level job.
---|---|---
Cambridge English Proficiency (CEP) | International executives, diplomats and others in humanities-oriented occupations | Certify the individual can communicate fluently in a second language (English) at a professional level
Test of English for International Communication (TOEFL) | International students | Certify the individual can communicate fluently in a second language (English) at a professional level
Test of English for International Communication (TOEIC) | International technical and business professionals | Certify the individual can communicate fluently in a second language (English) at a professional level
Tactical Iraqi/other Alelo simulation and game products | Military and business persons needing to engage with another culture | Prepare the individual for communicating in a foreign culture.

**Next Steps**

The vast majority of commercially available communications tests are straightforward language tests. They measure whether one’s language skills are sufficient to understand and to be understood. The implicit learning model and end goal are centered around giving/receiving information.

In the corporate world, the end goal is not always to provide information. Often it is to change the emotional state of the listener. Examples include calming an angry customer, reassuring a nervous mother, inspiring subordinates to give 110% in an all-hands meeting, engendering trust and warmth in a cold-called sales lead or an unwilling negotiating partner. Research is needed on how to achieve these more nuanced modes of communication, after which appropriate assessments can be developed. Sales professionals are an enormous potential market for tests designed to measure communication skills that persuade emotionally, as well as inform intellectually. Other promising directions include developing, evaluating, and iterating upon immersive communication tests which could test more nuanced situational communication skills and knowledge, and adapting global literacy assessments for employer use.

**Collaboration**

**Summary of Interest and the State of the Art**

Collaboration, and its closely related concept, teamwork, are consistently cited as major skill needs amongst employers.\(^{131}\)\(^{132}\)\(^{133}\)\(^{134}\) Yet there seems to be a gap at the nexus of theory, assessment, and validity.

There are teamwork assessments that work but whose core model has been invalidated, theoretical models of teamwork that have no accompanying assessment, teamwork assessments that have no proof/validity.

---

studies to back them up, and ad-hoc formulas derived from existing assessments that have predictive validity, but poor or no models.\textsuperscript{135} We were unable to locate the holy grail of a clear model, an assessment aligned with the model, and proof that the assessment works (correlates with supervisor/peer/objective ratings of teamwork performance or work product quality), in our search.

<table>
<thead>
<tr>
<th>Test Name</th>
<th>Deficiency</th>
<th>Overarching Workplace Purpose</th>
</tr>
</thead>
<tbody>
<tr>
<td>Stevens and Campion Teamwork KSA test</td>
<td>Poor psychometrics; appears to inadvertently measure cognitive ability</td>
<td>To predict individual’s performance in a team</td>
</tr>
<tr>
<td>Big 5 personality tests (various)</td>
<td>No individual-level model or individual-level teamwork predictions available.</td>
<td>Research on the applicability of personality traits to group-level work performance.</td>
</tr>
<tr>
<td>Composite test battery used by Morgeson (Morgeson, Reider, &amp; Campion, 2005)</td>
<td>An incredibly long and difficult assessment suite. Underlying model is conceptually muddy.</td>
<td>To predict individual’s performance in a team</td>
</tr>
<tr>
<td>ACT-WorkKeys Teamwork Test</td>
<td>Commercially available test but no documented validity studies.</td>
<td>To predict individual’s performance in a team</td>
</tr>
<tr>
<td>Kenexa Teamwork Indicator Assessment</td>
<td>No publicly available information other than the fact a trademark has been registered.</td>
<td>To predict individual’s performance in a team</td>
</tr>
<tr>
<td>After Action Review</td>
<td>Organizations outside the military succeed in having the structured discussion, but fail at committing to changed practices.</td>
<td>To improve team performance from one operation to the next.</td>
</tr>
</tbody>
</table>

\textit{Next Steps}

One approach might be to borrow assessments from the academic community, as there is a large need for a validated model and a validated commercial teamwork assessment. Another potential opportunity is to develop a genuine after-action review process, similar to what is used in the U.S. military, in which groups reflect on their collaborative work together in order to improve. This would function as a formative assessment with a focus on practice improvement.

\textsuperscript{135} For more detailed information, see the full report Workforce Assessments of 21st Century Competencies
The Character Dimension

Below is a summary table of the Advisory Panel’s findings on the current state of assessment of each character quality in the Four-Dimensional framework in the workforce.

<table>
<thead>
<tr>
<th>CCR Element</th>
<th>Importance to Employers</th>
<th>Most Commonly used assessment approaches in Workforce</th>
<th>Explanation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mindfulness</td>
<td>Low, but growing</td>
<td>Psychology self-assessments of the type used in medical and academic research.</td>
<td>Attention to Eastern mind-body practices such as yoga, transcendental meditation, etc. is increasing among employers, after a few CEOs championed the practices for themselves personally, and then also saw significant health, wellness, and productivity improvements among employees when the employees were offered mindfulness training as well. Corporate assessment of mindfulness is generally conducted in a research context (e.g., does this employee mindfulness program decrease employee absences?), using research tools.</td>
</tr>
<tr>
<td>Curiosity</td>
<td>Low</td>
<td>Specialized personality test, under development</td>
<td>A mixed relationships between curiosity and job performance has left employers not terribly interested in this construct. However, not much work has been done with this element, and it may be that higher level jobs require curiosity in larger measure than low-to-mid level jobs. Curiosity does appear to play a strong role in employees’ job satisfaction.</td>
</tr>
<tr>
<td>Courage</td>
<td>Low</td>
<td>Big 5 personality assessment</td>
<td>Not commonly assessed. Most prehire personality tests could easily be designed to look for courage (valid psychological test items for courage/bravery/valor exist in the public domain), but no major testing company that we could find, had such a category in its employer-side offerings. This implies a lack of market interest from employers with regard to testing for courage.</td>
</tr>
<tr>
<td>Resilience</td>
<td>High</td>
<td>Big 5 based personality assessment</td>
<td>Very commonly assessed via pre-hire personality tests, resilience is a known contributor to job performance in virtually all occupations.</td>
</tr>
<tr>
<td>Ethics</td>
<td>High</td>
<td>Self-report integrity questionnaires, using either overt or covert questions about counterproductive work behaviors (theft, dishonesty, attendance, verbal harassment, sexual harassment, sabotage, etc.)</td>
<td>Ethics is primarily assessed via a dedicated &quot;integrity test,&quot; taken during the job application process. Employers use these tests to look specifically for so-called counterproductive work behaviors: theft, fraud, dishonesty, sick leave abuse, cyberloafing, on-the-job substance abuse, verbal harassment, sexual harassment, sabotage, disregard for safety procedures, etc. Integrity testing is a mature industry with extensively validated tests. The assessments are popular with employers because the alternative is grim: theft alone costs a typical employer about 7% of its revenue every year.</td>
</tr>
<tr>
<td>Leadership</td>
<td>High, but situational</td>
<td>Assessment Centers, Structured Behavioral Interviews, Big 5 personality assessments</td>
<td>Corporate leadership tests are designed to assess whether an individual will perform well in a management position. They are therefore role-specific occupational assessments and not assessments of generic leadership ability. In addition to computerized tests, assessment centers - a physical space with trained observers - are used in testing leadership, one of the few instances in which their expense is justified.</td>
</tr>
</tbody>
</table>

The following sections go into more depth on each of these competencies, and suggest directions for future work.
Mindfulness

Summary of Interest and the State of the Art

Mindfulness interventions have been shown to reduce stress and stress-related illness (saving companies $2,000/yr per employee in health costs). Medical studies using objective task performance measures have shown mindfulness training will also increase workers’ task focus and short-term memory. These claims are corroborated by employee self-report surveys which document decreased stress, increased focus, and increased productivity after mindfulness training.

It is no surprise, then that corporations that have adopted mindfulness training as part of their efforts to improve their bottom line. As one example, Transport for London reports its mindfulness training workshops reduced employees’ overall absenteeism by 50%, and stress-related absenteeism by 70%.

Mindfulness assessments are currently used in a few corporations for research studies involving employees’ well-being and/or productivity. Consequently, research-grade psychology tools are the norm. The following assessments all have a self-report multiple choice format and have been well validated.

<table>
<thead>
<tr>
<th>Test Name</th>
<th>Distinguishing Feature</th>
<th>Overarching Workplace Purpose</th>
</tr>
</thead>
<tbody>
<tr>
<td>Five Facet Mindfulness Questionnaire (FFMQ)</td>
<td>5 subscales allows more in-depth analysis of an individual’s progression in mastering various aspects of mindfulness.</td>
<td>For research linking mindfulness levels in employees to productivity, stress, and health.</td>
</tr>
<tr>
<td>Mindful Attention and Awareness Scale (MAAS)</td>
<td>Simple daily life questions for probing individual’s general state of mindfulness.</td>
<td>For research linking mindfulness levels in employees to productivity, stress, and health.</td>
</tr>
<tr>
<td>12-item Cognitive and Affective Mindfulness Scale–Revised (CAMS-R)</td>
<td>Captures the flavor of the FFMQ but in a shorter instrument. Shorter length means no ability to extract different components of mindfulness.</td>
<td>For research linking mindfulness levels in employees to productivity, stress, and health.</td>
</tr>
</tbody>
</table>

Next Steps

Mindfulness is trendy and new at the moment, but it has the potential to become more widely accepted and increasingly well quantified (via assessments), if it continues to meet the very real corporate needs of keeping healthcare costs down, reducing absenteeism, and improving worker productivity.

A key gap in the mindfulness work is any quantitative research linking corporate leaders’ or general workers’ mindfulness states or practices to the improved bottom lines of their companies. Conferences on mindfulness and leadership abound as do consulting companies ready to show corporate leaders how to become more mindful. However, other than personal anecdotes from a handful of visible CEO

---

champions, a substantive link between mindfulness at the top, and profits at the bottom line, has yet to be established.

Curiosity

Summary of Interest and the State of the Art
Curiosity has been found to be correlated with level of sophistication of one’s occupation, as well as peer and supervisor ratings of job performance, as well as life and job satisfaction, although not all the research has shown strong effects. Using the Values in Action (VIA) Character Strengths survey, Peterson and coworkers have shown high correlations between curiosity and job satisfaction, with the correlation strongest in homemakers and professionals, and lowest (though still notable) in blue collar and clerical workers.

At least one assessment has been developed to assess curiosity in the workforce explicitly; however it is still in the research phase of development. For most of the Big 5 based assessments currently used in pre-hire selection, curiosity is an extremely minor or non-contributing component because the dimension in which it is based, Openness, is the least important Big 5 dimension to job performance in most jobs. It appears far more prominently in assessments used primarily for personal growth, such as the VIA assessment.

<table>
<thead>
<tr>
<th>Test Name</th>
<th>Methods</th>
<th>Overarching Workplace Purpose</th>
</tr>
</thead>
<tbody>
<tr>
<td>German Work-Related Curiosity Scale</td>
<td>Likert Scale (Multiple Choice)</td>
<td>Research to determine whether curiosity impacted work performance</td>
</tr>
<tr>
<td>VIA</td>
<td>Likert Scale (Multiple Choice)</td>
<td>Research to determine whether curiosity impacted job satisfaction</td>
</tr>
<tr>
<td>Hogan Select</td>
<td>Likert Scale (Multiple Choice)</td>
<td>Pre-hire selection. Curiosity is a minor component of Hogan's Dependability scale.</td>
</tr>
<tr>
<td>Knack (Wasabi Waiter, MetaMaze)</td>
<td>Game</td>
<td>Career and self exploration</td>
</tr>
</tbody>
</table>

141 "sophistication" = a number equivalent to the job’s O’Net Job Zone category
Next Steps

There appears to be a need for basic research on the role of curiosity in the workforce. The German Work Related Curiosity Scale found curiosity demonstrated a moderately strong correlation to one’s progression up the occupational ladder, even though its relationship to job performance was mixed.\(^\text{152}\) This suggests that curiosity may be important to the more sparsely populated, higher level occupations typically ignored by Big 5-based pre-hire screening tests.

However, there appears to be no work (as yet) identifying exactly which occupations these might be. The existing research with the German Work Related Curiosity Scale aggregates multiple occupations into a single analysis.\(^\text{153, 154}\) Thus, one area of future work is to take occupations that might be likely candidates – those requiring in-depth analysis or wide-ranging exploration – and determine the importance of curiosity to these specific occupations, thereby establishing an as-yet-unused screening dimension for higher level work.

In addition, it would seem that curiosity is ripe for a game-based assessment dedicated primarily to uncovering this trait. The two components of curiosity, exploration and absorption, are naturally accessible from gameplay behaviors. Recording how many new chapters, characters, or lands a gamer explores and experiments with, is likely to be a valid measure of exploration or interest-driven curiosity. Monitoring how long and deeply a gamer delves into each new challenge or experience would be a measure of absorption-based curiosity. Hypothetically, user data from existing turn-based strategy games or expansive virtual worlds (games in which it is possible to spend a lot of time exploring the content and delving into strategy complexity) could be resold to firms looking for high curiosity-driven individuals to recruit.

Courage

**Summary of Interest and the State of the Art**

One individual-level\(^\text{155, 156}\) and one corporate-level\(^\text{157, 158}\) assessment of courage have been used for research purposes. Personal bravery contributes to an employee’s job dedication and organizational support,\(^\text{159}\) and is positively related to subordinates’ ratings of an executive’s job performance.\(^\text{160}\)

---


\(^{156}\) International Personality Item Pool, "The Items Included in Each of the 24 Revised IPIP-VIA Scales," 2015. [Online].

\(^{157}\) International Personality Item Pool, "The Items Included in Each of the 24 Revised IPIP-VIA Scales," 2015. [Online].


Organizational bravery results that correlate to personal reports of workplace satisfaction and burnout or lack thereof.\textsuperscript{161} While both assessments are readily available, neither appears to be used for typical corporate end uses such as employee screening, training, or development. This may be because courage is difficult to define, and appears to be very context-specific.

<table>
<thead>
<tr>
<th>Test Name</th>
<th>Methods</th>
<th>Overarching Workplace Purpose of Test</th>
</tr>
</thead>
<tbody>
<tr>
<td>VIA (Values in Action)</td>
<td>Likert Scale/Multiple Choice</td>
<td>Research to determine whether being courageous impacts personal job satisfaction or performance</td>
</tr>
<tr>
<td>Kilmann Organizational Courage</td>
<td>Likert Scale/Multiple Choice</td>
<td>Corporate self-exploration. Also, research on whether a courageous corporate climate impacts employee satisfaction.</td>
</tr>
</tbody>
</table>

\textit{Next Steps}

Based on the limited evidence available, one appropriate next step may be to combine personality test assessments with a corporate climate assessment, then try to develop a courage model containing the results of these two surveys to predict the frequency with which an individual in an organization is likely to commit a “qualified brave action” within a specified time period. Gathering data on the actual frequency of brave workplace actions would then allow the model – and by extension, the measures contributing to the model – to be validated. It would also be important to consider and analyze observable differences across contexts and include this in the learning model. Another possible direction is to validate the organizational courage assessment against external performance measures such as sales, stock growth, employee turnover, employee promotion rates, etc.

\textbf{Resilience}

\textit{Summary of Interest and the State of the Art}

Resilience is a desirable trait in the workforce, as it describes the ability to recover from stressful incidents, and thus the ability to avoid stress related health problems and absences, emotional outbursts, defensive behaviors, moodiness and unpredictability.\textsuperscript{162,163} Resilience and its synonyms (see table below) bear notable ($r>0.2$) correlations with observer ratings of workers’ emotional maturity, ability to persuade others, negotiation skill, entrepreneurial acumen, maintaining optimism, showing interpersonal understanding, showing concern for quality, demonstrating business acumen, and effectiveness in delegating and monitoring assignments.\textsuperscript{164,165} Not surprisingly, resilience is positively correlated to job performance in virtually every occupation tested.\textsuperscript{166,167}

\begin{itemize}
  \item \textsuperscript{162} Hogan Assessment Systems, \textit{Hogan Select Technical Manual}. Tulsa, OK, 2009.
  \item \textsuperscript{163} J Russo et al., "Neuroticism and extraversion as predictors of health outcomes in depressed primary care patients.." vol. 38, pp. 339-48, Jul-Aug 1997.
  \item \textsuperscript{166} Ibid.
  \item \textsuperscript{167} Ibid.
\end{itemize}
Resilience is predominantly assessed via the neuroticism (emotional stability) dimension of Big 5-based personality tests. Using Likert scales, individuals respond to self-report questions as, “seldom feel blue,” “rarely get irritated,” “am not easily bothered by things,” “remain calm under pressure”.\(^{168}\) While the trait is called different things by different test vendors, it is possible to tell whether a given trait is resilience or not, by whether it falls out in the neuroticism category during factor analysis of the test results, or by comparing sample test questions to those listed above. Many of the vendors offering Big 5-based personality tests feature resilience-specific data report-outs; a few such tests are summarized below, along with what they call this trait.

<table>
<thead>
<tr>
<th>Test Name</th>
<th>Name of Trait Corresponding to Resilience</th>
<th>Overarching Workplace Purpose of Test</th>
</tr>
</thead>
<tbody>
<tr>
<td>Kanexa/IBM, Leadership Preference Assessment</td>
<td>Emotional Orientation</td>
<td>To help identify candidates for promotion to leadership positions.</td>
</tr>
<tr>
<td>Hogan Select, Hogan Personality Inventory</td>
<td>Composure, Adjustment</td>
<td>Pre-hire selection. This trait is included for suitability matching against virtually all job families.</td>
</tr>
<tr>
<td>SHL Occupational Personality Questionnaire</td>
<td>Coping with Pressures and Setbacks</td>
<td>Pre-hire selection, promotion and/or re-assignment of incumbents</td>
</tr>
<tr>
<td>Caliper Profile</td>
<td>Ego Strength/Resilience</td>
<td>Pre-hire selection; this trait is included for suitability matching against virtually all job families.</td>
</tr>
</tbody>
</table>

**Next Steps**

While resilience is clearly a priority for the workforce and is assessed validly and reliably, it may be useful to continue to research the development of resilience and contributing factors, in order to create a learning model that would help institutions improve, not just assess, resilience.

**Ethics**

**Summary of Interest and the State of the Art**

Because of its impact on the bottom line, employers take ethics testing quite seriously, as it is estimated that globally 5% of annual corporate revenues are disappearing due to fraud.\(^{169}\) Of all the CCR elements, ethics seems to have the best developed workforce assessments. Paul Barrett’s recent summary of existing commercial integrity assessments lists 26 current commercial integrity tests alone,\(^{170}\) along with detailed instructions on how to evaluate different assessments for purchase.\(^{171}\) These tests fall into two categories: overt and covert.

In overt tests, respondents respond to statements about Counterproductive Work Behaviors (CWB’s) such as “I have stolen merchandise within the past 3 years” (choose yes or no). Surprisingly, (either because

\(^{169}\) Association of Certified Fraud Examiners, "Report to the Nations on Occupational Fraud and Abuse," ACFE, Austin, TX, 2014.
the respondents really believe what they are doing is right,\textsuperscript{172} \textsuperscript{173} \textsuperscript{174} or because they interpret an “integrity test” as one in which they will be judged on the honesty of their answers\textsuperscript{175}) these tests produce valid results. It has been shown, however, that if instructed to fake responses, respondents can increase their scores by almost a full standard deviation.\textsuperscript{176} \textsuperscript{177} In one variation of an overt test, the Conditional Reasoning Test of Aggression, respondents choose between a pair of responses that reflect a sunny, open interpretation or a dark, paranoid interpretation of the same set of facts. This test has been consistently shown to be highly valid.\textsuperscript{178}

Covert tests are personality tests designed to detect individuals high in conscientiousness, high in agreeableness, and low in neuroticism, as such people tend to be less prone to CWB’s.\textsuperscript{179} While less easy to fake than overt tests, the covert tests are often less valid, because of the uncertainty involved in projecting actual behavior from personality (tendency to behavior). In contrast, overt tests predict future behavior from past behavior. Ethics can also validly be detected in structural behavioral interviews.\textsuperscript{180}

<table>
<thead>
<tr>
<th>Test Name</th>
<th>Methods</th>
<th>Overarching Workplace Purpose</th>
</tr>
</thead>
<tbody>
<tr>
<td>Personnel Selection Inventory</td>
<td>Overt Integrity Test (True/False)</td>
<td>Predict counterproductive work behaviors (CWB’s) of job applicants</td>
</tr>
<tr>
<td>Reid Report</td>
<td>Overt Integrity Test (True/False)</td>
<td>Predict counterproductive work behaviors (CWB’s) of job applicants</td>
</tr>
<tr>
<td>Stanton Survey</td>
<td>Overt Integrity Test (True/False; some Multiple Choice)</td>
<td>Predict counterproductive work behaviors (CWB’s) of job applicants</td>
</tr>
<tr>
<td>Certifinder Applicant Review</td>
<td>Overt Integrity Test (format not disclosed)</td>
<td>Predict counterproductive work behaviors (CWB’s) of job applicants</td>
</tr>
<tr>
<td>Conditional Reasoning Test of Aggression (CRT-A)</td>
<td>Overt Integrity Test (Multiple Choice)</td>
<td>Predict counterproductive work behaviors (CWB’s) of job applicants</td>
</tr>
<tr>
<td>California Psychological Inventory</td>
<td>Covert Integrity Test (True/False)</td>
<td>Predict overall job performance of job applicants, of which counterproductive work behaviors (CWB’s) are a part</td>
</tr>
<tr>
<td>The Personnel Reaction Blank</td>
<td>Covert Integrity Test (Multiple Choice)</td>
<td>Predict counterproductive work behaviors (CWB’s) of job applicants</td>
</tr>
<tr>
<td>Giotto</td>
<td>Covert Integrity Test (Forced Choice)</td>
<td>Predict counterproductive work behaviors (CWB’s) of job applicants</td>
</tr>
<tr>
<td>Inwald personality Inventory-2</td>
<td>Covert Integrity Test (True/False)</td>
<td>Predict counterproductive work behaviors (CWB’s) of job applicants</td>
</tr>
<tr>
<td>Hogan Personality Inventory</td>
<td>Covert Integrity Test (True/False)</td>
<td>Predict overall job performance of job applicants, of which counterproductive work behaviors (CWB’s) are a part</td>
</tr>
</tbody>
</table>

\textsuperscript{179} Deniz S. Ones, "Establishing Construct Validity for Integrity Tests," University of Iowa, Iowa City, IA, DTIC 19950505173 1993.
Structured Behavioral Interviews

<table>
<thead>
<tr>
<th>Structured Behavioral Interviews</th>
<th>In-person interview with questions designed to elicit specific examples of ethical/unethical behavior from the interviewee’s past. Answers evaluated by an evaluation team via a rigorous rubric.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Predict overall job performance of job applicants, of which counterproductive work behaviors (CWB’s) are a part</td>
<td></td>
</tr>
</tbody>
</table>

**Next Steps**

Looking forward, the paper and pencil integrity test may someday be augmented by physiological measures. Conventional lie detectors are neither valid nor legally admissible, however, experimentation with facial recognition technology has produced facial action coding software capable of detecting many of the facial tics that accompany lying. If legally accepted, this tool would allow interviewers to ask questions about an individual’s past and determine more accurately whether the answers given are honest ones.

In the medical research arena, MRI scans now provide insight into the mental structures associated with lying and other forms of deviant behavior. These discoveries could potentially feed into redesigned paper and pencil tests where specific questions are designed to trigger specific brain pathways. At present, the structure of ethics tests (for the covert tests, at least) is based on hypothetical factors that exist only in the data patterns of assessment answer sheets.

**Leadership**

**Summary of Interest and the State of the Art**

At present, most corporate assessments for leadership are testing specifically for management skills. In assessment centers, the assessment tasks are managerial tasks. In structured behavioral interviews, the questions are designed around managerial tasks extracted from a job analysis. Candidates are asked for examples from their own past that demonstrate they have successfully overcome typical challenges associated with those tasks. In personality testing, an overview of the individual’s personality is obtained, but when that profile is compared to the boundaries of “acceptable” personalities for leadership roles, it is in fact comparing the candidate to high performing (or even standard performing) corporate managers.

Of the three approaches to determining the best future manager, the structured behavioral interview has the strongest predictive validity, as well as very high racial equity (compared to unstructured interviews, racial bias is decreased 50-100%). However, the methodology is time-consuming and therefore best used when the position to be filled is of critical importance (so that the extra effort is warranted), or when there will be multiple hires into the same job title (in which case the effort is amortized amongst the multiple hires). The assessment center is the next best choice, in terms of predictive validity and ability to

---

http://mplab.ucsd.edu/grants/project1/research/Fully-Auto-FACS-Coding.html


custom match the assessment to job tasks. The least accurate and least expensive choice – though still adequate – is mass screening via cognitive and/or psychological multiple choice test items.

One assessment that exists largely in the academic sphere, but has displayed astonishing (perhaps unbelievable) psychometrics when deployed in one study with corporate managers is the Student Leadership Practices Inventory. What sets this assessment apart are two attributes. First, its five categories of questions adhere much more closely to the sense of “leadership” found in leadership books: “model the way,” “inspire a shared vision,” “challenge the process,” “enable others to act,” and “encourage the heart.” These vague-sounding categories are demonstrably real and distinct: they separate out as 5 separate concepts in factor analysis of respondents’ answers to the questions in the survey. Secondly, the ability of this instrument to provide a consistent framework for leadership ability is compelling. In a study with 708 managers, subordinates’ ratings of their managers’ overall performance could be correctly predicted to R=0.8 from component ratings (by the same subordinates) on the 5 pillars. While this is not as valuable as having a self-report that generates the same validity, the consistency of the framework suggests it would be fruitful to evaluate the validity of a self-report version of the same instrument (which conveniently does already exist).

<table>
<thead>
<tr>
<th>Test Name</th>
<th>Methods</th>
<th>Overarching Workplace Purpose</th>
</tr>
</thead>
<tbody>
<tr>
<td>DDI Assessment Center</td>
<td>Physical space with simulated management tasks and trained observers</td>
<td>Pre-hire selection - To rate individual's generic management potential</td>
</tr>
<tr>
<td>DDI People Leader</td>
<td>Web-based tasks with remote evaluators assessing candidate's behavior during task performance.</td>
<td>Pre-hire selection - To rate individual's generic management potential</td>
</tr>
<tr>
<td>OPM Structured Behavioral Interview</td>
<td>In-person interview with questions designed to elicit specific job task performance examples from the interviewee's past. Answers evaluated by an evaluation team via a rigorous rubric.</td>
<td>Pre-hire selection - To rate individual's match to a specific job opening</td>
</tr>
<tr>
<td>Berke Assessment</td>
<td>Online multiple choice test containing cognitive items (e.g., vocabulary, spatial visualization) and personality items</td>
<td>Pre-hire selection - To rate individual's generic management potential</td>
</tr>
<tr>
<td>Hogan Select – High Potential</td>
<td>Multiple choice tests with self-reports of personality traits</td>
<td>Pre-hire selection - To rate individual's generic management potential</td>
</tr>
<tr>
<td>HoganLead</td>
<td>Multiple choice tests with self-reports of personality traits</td>
<td>Self-development</td>
</tr>
<tr>
<td>SHL Occupational Preference Questionnaire</td>
<td>Multiple choice tests with self-reports of personality traits</td>
<td>Pre-hire selection - To rate individual's generic management potential</td>
</tr>
<tr>
<td>Leadership Practices Inventory</td>
<td>Web-based multiple choice self-report that can be corroborated with 360 reports from others.</td>
<td>Research and self-development</td>
</tr>
</tbody>
</table>

**Next Steps**

---

Overall, the direction for growth in this area is to assess Leadership rather than management. None of the existing assessments was designed to detect a future Ghandi, Nelson Mandela, George Washington. And, while many popular books have been written about leadership, those books do not come paired with validated assessments to make their motivational concepts real and measurable. Thus, we have a set of management assessments that is not designed to predict leadership in any other sphere, and a set of inspirational leadership treatises that do not have a corresponding assessment.

Of the existing management assessments, structured behavioral interviews have been shown to be highly valid and racially equitable. However, they are time consuming, due to the time requirements of the in-person interview.

The self-report Leadership Practices questionnaire, if it can be validated in self-report form, promises an entirely different underlying paradigm – one not based on personality or tasks – but on conceptual subdivisions of leadership.
The Meta-Learning Dimension

The ability to “learn how to learn” and transfer what has been learned from one context to the next, is a very popular topic among employers. In the workforce this has been coined “learning agility,” corresponding most closely with the entire dimension of Meta-Learning, rather than either of its elements. Metacognition seems to be completely missing from workforce discussions. For this reason, this report will focus on learning agility and growth mindset in the workforce, to approximate the Meta-Learning dimension.

Below is a summary table of the Advisory Panel’s findings on the current state of assessment of the Meta-Learning dimension of the Four-Dimensional Framework.

<table>
<thead>
<tr>
<th>CCR Element</th>
<th>Importance to Employers</th>
<th>Most Commonly used Assessment Methods in Workforce</th>
<th>Explanation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Metacognition</td>
<td>None</td>
<td>Two assessments of “learning agility”: one is multi-rater (and used to assess personal or executive development) and the other is self-report (used to screen candidates for managerial positions).</td>
<td>No instruments measuring metacognition in the workplace could be found to date. However, learning agility is a separate concept used to explain the same end result - namely the ability of some individuals (workers) to come up to speed quickly in new situations, or to be able transfer learning from one context to the next. The concept of learning agility is wildly popular in corporate leadership blogs and has two well-known assessment tools associated with it.</td>
</tr>
<tr>
<td>Growth Mindset</td>
<td>None</td>
<td></td>
<td>Other than a single research study involving Carol Dweck and the Dweck instrument, there has been no visible work assessing growth mindset in corporations.</td>
</tr>
</tbody>
</table>

Learning Agility  
(alternative to Metacognition)

Summary of Interest and the State of the Art

---

The Center for Creative Leadership (CCL) defines learning agility as the ability to “give up skills, perspectives, and ideas that are no longer relevant, and learn new ones that are.”\(^{196}\) Employers tend to have a personality-oriented view of learning agility, rather than viewing it as comprising practices such as Metacognition and Growth Mindset. In other words, employers view the primary obstacle to learning as simply being open to learning in the first place.

The iconic learning agility assessment is the Choices™ tool, which attempts to assess four supporting constructs that are presumed to contribute to learning agility: mental agility, people agility, change agility, and results agility.\(^{197}\) The consulting firm Korn-Ferry adopted and revamped the Choices™ tool, adding a fifth concept, self-awareness, to result in a final suite of tested concepts as listed below:\(^{198}\)

- **Mental agility**—comfortable with complexity
- **People agility**—skilled communicator who can work with a diversity of people
- **Change agility**—like to experiment and comfortable with change
- **Results agility**—deliver results in first-time situations
- **Self-awareness**—the depth to which an individual recognizes skills, strengths, weaknesses, blind spots, and hidden strengths

The Korn-Ferry toolset seems to be the primary commercially available tool for learning agility at present.

Initial research has shown that scores on the original Choices™ tool predict a supervisor’s opinion of whether an individual should be promoted \((r=0.4)\), and supervisors’ opinions on that person’s job performance \((r=0.37)\).\(^{199}\) These validity levels are unusually strong for a psychology-based instrument. The consulting firm Korn-Ferry conducted some subsequent internal studies showing that executives who scored as “highly agile” on Korn-Ferry’s learning agility instrument had 25% higher profit margins than their peer group,\(^ {200}\) were promoted twice as fast\(^{201}\)\(^ {202}\), and were 5 times more likely to be highly engaged.\(^ {203}\)

There are, however, reasons to be cautious about these results. The Choices™ instrument contains a few items that directly ask about performance, rather than agility per se (meaning its score would automatically correlate with other instruments that ask about performance)\(^ {204}\) and it was designed as a 360° assessment, so the strong correlations may be due more to the same person (typically a supervisor) filling out both the learning agility questionnaire and the job performance questionnaire, than a genuine

---

relationship between learning agility and on-the job performance. Due to the limited amount of validation work in existence, there is not yet an expert consensus on the predictive ability of learning agility assessments.

Although learning agility is promoted as a personality construct, it is not clear that it is one, since only two of the Big 5 dimensions (Openness and Conscientiousness) had even a small correlation to learning agility scores. There was also no correlation to cognitive skill, and high learning agility individuals generally are not detail oriented, planful, or methodical – in stark contradiction to individuals high in metacognition.

<table>
<thead>
<tr>
<th>Test Name</th>
<th>Test Features</th>
<th>Overarching Workplace Purpose</th>
</tr>
</thead>
<tbody>
<tr>
<td>Choices™, Choices Architect</td>
<td>Multi-rater tool with 81 items</td>
<td>Personal or executive development</td>
</tr>
<tr>
<td>Via Edge</td>
<td>Self-report tool with 116 items, including items designed to detect faking</td>
<td>Screening candidates for managerial positions</td>
</tr>
</tbody>
</table>

**Next Steps**

Since Metacognition is a concept not explored by the workforce, it may be interesting to do a pioneering study involving its role in various jobs. For example, it may be interesting to correlate level of metacognition to learning agility, and to various other outcome measures such as job sophistication, job satisfaction, and job performance.

Given the dichotomous views between academia and employers on what makes for a fast, efficient, and adaptive learner (metacognition vs. learning agility), it would be fascinating to place these two concepts in a head-to-head competition. The same individuals could be given assessments for both metacognition and learning agility, then given unfamiliar tasks to master. One set of tasks could be more academic in nature; the other set of tasks could be more job-oriented. Task performance (using a quality per unit time metric) would serve as the final arbiter of the ability to learn new things quickly. Back-correlating to the assessment scores for metacognition and learning agility would allow researchers to see which concept predicted new task mastery the best -- or at least isolate which kinds of tasks each concept was most relevant to.

Also, since the personality-oriented view seems to reflect a fixed mindset orientation; it would be interesting to assess employers’ views on the innateness of this attribute, and whether it is affecting their subordinates’ development in learning agility; if this was true, there would be a positive correlation between the level of employers’ Growth Mindset approach, and their subordinates’ scores.

**Growth Mindset**

**Summary of Interest and the State of the Art**


The Advisory Panel located one research study using a growth mindset instrument in a corporate setting. Specifically, a variation on Carol Dweck’s famous tool was used by Dweck and coworkers to evaluate whether corporate cultures have a growth or fixed mindset. In the corporate version of the test, the underlying growth vs. fixed mindset dichotomy was re-cast as a “culture of development” vs. a “culture of genius” for the company as a whole. The results indicated corporations with a culture of development engendered higher levels of employee trust and ownership, led employees to be more innovative and collaborative, and lowered unethical behavior.208

Aside from the one research study, however, there is little on growth mindset in the corporate testing literature. Dweck has argued that those with a growth mindset will exhibit superior perseverance through obstacles or tasks, compared to those with a fixed mindset.209 If this is so, then perseverance traits could serve as a proxy for growth mindset. As it turns out, perseverance is a facet of the Big 5 personality dimension, Conscientiousness.210 Thus, the supposed outcome of a growth mindset is already being probed via any one of the many commercially available Big 5-based personality tests. However, the Big 5-based assessments are framed assuming perseverance is the consequence of a personality trait rather than the outcome of a mental belief system. The strongly agree <--> strongly disagree Likert scale questions in the Big 5 assessments are along the lines of “I complete tasks successfully”211 (this is who I am) rather than the Dweck assessment’s “to be honest, you can’t really change how intelligent you are”212 (this is what I believe).

Next Steps
Growth mindset related assessments do not appear to be utilized in the workplace at this time, but one could argue that they be adopted. In order to be useful to human resource officers, a growth mindset assessment would have to outperform existing assessments for pre-hire selection. Specifically, the growth mindset assessment score would need to be positively correlated to some aspect of job performance at a level r>0.2 (a typical correlation between Big 5 personality traits and job performance213) and, moreover, the assessment scores would need to be racially equitable.

Alternatively, one could target Chief Learning Officers as potential adopters – perhaps using growth mindset assessments to diagnose slow learning progress during in-house training or identify the candidates most likely to benefit from tuition reimbursement programs. However, even in this case, one would need stronger validity evidence. The correlation of growth mindset scores to learning gains and expended effort is solid214 215, but their correlation to absolute academic performance (GPA, standardized

---

test scores) is far weaker and often negative.\textsuperscript{216} \textsuperscript{217} \textsuperscript{218} This is because an individual’s final level of academic achievement depends not only on the interim gain, but on their pre-existing level of achievement, which in turn depends far more strongly on variables other than effort, such as socioeconomic status.\textsuperscript{219} \textsuperscript{220} 

\begin{itemize}
  \item \textsuperscript{216} Ann E. Dvorak, "Student Mindset Compared to Performance on the Nebraska State Accountability Test," \textit{DigitalCommons@University of Nebraska - Lincoln}, 2014. [Online]. http://digitalcommons.unl.edu/cgi/viewcontent.cgi?article=1101&context=aglecdiss
  \item \textsuperscript{218} Adrian Furnham, Tomas Chamorro-Premuzic, and Fiona McDougall, "Personality, cognitive ability, and beliefs about intelligence as predictors of performance," \textit{Learning and Individual Differences}, vol. 14, pp. 47-84, 2002.
  \item \textsuperscript{220} Chris Cook, "London’s Big Lead and Social Mobility," \textit{Financial Times}, August 5 2012. [Online]. http://blogs.ft.com/ftdata/2012/08/05/londons-big-lead-and-social-mobility/\
\end{itemize}
Summary Table for Workforce Assessments:

<table>
<thead>
<tr>
<th>CCR Element</th>
<th>Importance to Employers</th>
<th>Most Commonly used Assessment Methods in Workforce</th>
<th>Explanation</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Skills</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Creativity</td>
<td>Low to Negative (high in leadership)</td>
<td>Game, Make-a-List Idea Fluency Exercises, Big 5 personality assessment (Likert scale/multiple choice self-report)</td>
<td>Ranked dead last by employers on a major survey on what's important to them in hiring. Also, given that pre-hire personality tests screen for conscientiousness, they will automatically screen against creativity, since these two personality traits are anticorrelated in most of the population. On the other hand, it has been rated as the #1 factor necessary for future success in a global CEO study. The discrepancy might be explained by different characteristics being desired for different levels in the hierarchy.</td>
</tr>
<tr>
<td>Critical Thinking</td>
<td>High</td>
<td>Situational judgment tests such as the Watson-Glaser. However, this test – and others like it – is rarely used at the present time.</td>
<td>Ranked in the top 3 of requested skills by major employer surveys. Disproportionately little testing of this construct in the US due to legal challenges arising from large score differentials between majority and minority ethnic/racial groups.</td>
</tr>
<tr>
<td>Communication</td>
<td>High</td>
<td>Degree (and personal interview, cover letter, writing sample, etc.) is most common. Reading, Writing, Listening and Speaking tests are less common.</td>
<td>Also a &quot;top 3&quot; skill in employer surveys, communication is predominantly assessed by proxy - assuming that a college or high school degree equates to an acceptable level of reading, writing, and speaking skill. Secondarily, these skills are assessed via informal in-person means: review of resume and cover letter (to judge writing skill), job interview conversation (gives some indication of oral communication skill), and response to written instructions during the application process (reading skills). A few niche markets, e.g., employees needing to conduct international businesses, take part in explicit and rigorous communications testing.</td>
</tr>
<tr>
<td>Collaboration</td>
<td>High</td>
<td>Stevens-Campion Teamwork KSA test and/or individual personality traits the employer feels probably relate to teamwork, from Big 5 personality assessments</td>
<td>The last of the &quot;top 3&quot; in-demand skills from employer surveys, this skill has a bewildering array of conceptual models available, but little in the way of workplace-validated assessments.</td>
</tr>
<tr>
<td><strong>Character</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mindfulness</td>
<td>Low, but growing</td>
<td>Psychology self-assessments of the type used in medical and academic research.</td>
<td>Attention to Eastern mind-body practices such as yoga, transcendental meditation, etc. is increasing among employers, after a few CEOs championed the practices for themselves personally, and then also saw significant health, wellness, and productivity improvements among employees when the employees were offered mindfulness training as well. Corporate assessment of mindfulness is generally conducted in a research context (e.g., does this employee mindfulness program decrease employee absences?), using research tools.</td>
</tr>
<tr>
<td>Curiosity</td>
<td>Low</td>
<td>Specialized personality test, under development</td>
<td>A mixed relationship between curiosity and job performance has left employers not terribly interested in this construct. However, not much work has been done with this element, and it may be that higher level jobs require curiosity in larger measure than low-to-mid level jobs. Curiosity does appear to play a strong role in employees' job satisfaction.</td>
</tr>
<tr>
<td>Courage</td>
<td>Low</td>
<td>Big 5 based personality assessment</td>
<td>Not commonly assessed. Most prehire personality tests could easily be designed to look for courage (valid psychological test items for courage/bravery/valor exist in the public domain), but no major testing company that we could find, had such a category in</td>
</tr>
</tbody>
</table>
Its employer-side offerings. This implies a lack of market interest from employers with regard to testing for courage.

<table>
<thead>
<tr>
<th>Resilience</th>
<th>High</th>
<th>Big 5 based personality assessment</th>
<th>Very commonly assessed via pre-hire personality tests, resilience is a known contributor to job performance in virtually all occupations.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ethics</td>
<td>High</td>
<td>Self-report integrity questionnaires, using either overt or covert questions about counterproductive work behaviors (theft, dishonesty, attendance, verbal harassment, sexual harassment, sabotage, etc.)</td>
<td>Ethics is primarily assessed via a dedicated &quot;integrity test,&quot; taken during the job application process. Employers use these tests to look specifically for so-called counterproductive work behaviors: theft, fraud, dishonesty, sick leave abuse, cyberloafing, on-the-job substance abuse, verbal harassment, sexual harassment, sabotage, disregard for safety procedures, etc. Integrity testing is a mature industry with extensively validated tests. The assessments are popular with employers because the alternative is grim: theft alone costs a typical employer about 7% of its revenue every year.</td>
</tr>
<tr>
<td>Leadership</td>
<td>High, but situational</td>
<td>Assessment Centers, Structured Behavioral Interviews, Big 5 personality assessments</td>
<td>Corporate leadership tests are designed to assess whether an individual will perform well in a management position. They are therefore role-specific occupational assessments and not assessments of generic leadership ability. In addition to computerized tests, assessment centers - a physical space with trained observers - are used in testing leadership, one of the few instances in which their expense is justified.</td>
</tr>
</tbody>
</table>

**Meta-Learning**

| Metacognition       | None       | No metacognition assessment. Instead, two assessments of "learning agility": one is multi-rater (and used to assess personal or executive development) and the other is self-report (used to screen candidates for managerial positions). | No instruments measuring metacognition in the workplace could be found to date. However, learning agility is a separate concept used to explain the same end result - namely the ability of some individuals (workers) to come up to speed quickly in new situations, or to be able transfer learning from one context to the next. |
| Growth Mindset      | None       | Other than a single research study involving Carol Dweck and the Dweck instrument, there has been no visible work assessing growth mindset in corporations. | |

---

Evolving Assessments for a 21st Century Education © Center for Curriculum Redesign – All Rights Reserved Page 92
VIII. RECOMMENDATIONS
Possible Next Steps for the ARC (priorities and sequence TBD by ARC Members)

The following recommendations are to be considered a “starter set” derived from the findings and results of the initial review work of the Advisory Panel. They are by no means comprehensive, complete or detailed enough to immediately start implementing the recommendations. ARC Members will need to prioritize, sequence and add further detail on how these, and recommendations that Members will bring to the table, can be effectively implemented over the many future phases of the Consortium work.

A. GENERAL ASSESSMENT RECOMMENDATIONS

1. Research Learning Model Competencies and Learning Progressions
One of the biggest challenges facing the ARC Consortium is to launch a research agenda to create useful learning competency models and progressions for each of the Knowledge, Skills, Character and Meta-Learning elements and sub-elements in the assessment framework, as suggested by the highlighted base of the “circled triangle” assessment model presented earlier:

![Model of Assessment Diagram]

This work will have to be done in phases, with ARC Members deciding on the priority and sequence of framework elements to be researched.

Though some areas of Traditional Knowledge (Reading, Maths and Science) and some Skills have well-researched competency models and learning progressions, very little exists for other Knowledge areas (especially Modern Knowledge competencies) and Skills development, and even less for Character qualities and Meta-Learning strategies.
This work done on researching and creating learner competency development models will directly feed into supporting a research-based process for developing new assessments.

### 2. Further Research the Alignments and Gaps Between Existing Assessment Methods and the Various Purposes of Assessments

To create a robust framework for 21st century assessment research and practice, one of the early steps might be to do further research on existing alignments between the purposes and methods of assessment for each of the elements in the CCR Four-Dimensional Education framework, and develop guidance from research on which methods are more (or less) effective in supporting each purpose for each dimensional element (and its sub-elements).

The challenge will be to do the appropriate research to fill in this Assessment Guidance chart and to identify the assessment gaps that need further development:

<table>
<thead>
<tr>
<th>Framework Dimensions &amp; Elements</th>
<th>Methods for Individual Diagnosis</th>
<th>Methods for Practice Improvement</th>
<th>Methods for Accountability</th>
<th>Methods for Program Evaluation</th>
<th>Methods for Research</th>
</tr>
</thead>
<tbody>
<tr>
<td>KNOWLEDGE</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(Elements TBD)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SKILLS</td>
<td>Creativity</td>
<td>Critical Thinking</td>
<td>Communication</td>
<td>Collaboration</td>
<td></td>
</tr>
<tr>
<td>CHARACTER</td>
<td>Mindfulness</td>
<td>Curiosity</td>
<td>Courage</td>
<td>Resilience</td>
<td>Ethics</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Leadership</td>
</tr>
<tr>
<td>META-LEARNING</td>
<td>Metacognition</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

© Center for Curriculum Redesign – All Rights Reserved
Once complete, this chart will still not ensure the appropriate contextual use of assessments, but it can serve as a general “rule-of-thumb” guide to assessment areas needing more attention.

3. **Adopt/Adapt/Create a Research-based Process for Developing New Assessments**

It is clear from the Advisory Panel’s surveys of existing assessments that new assessments will need to be created to effectively assess all the elements in the Four-Dimensional Education framework. One such well-developed methodology (especially relevant for developing valid assessments for 21st century competencies) is the Evidence Centered Design approach, pioneered by researchers at ETS and championed by Robert Mislevy who states:

ECD provides a conceptual design framework for the elements of a coherent assessment at a level of generality that supports a broad range of assessment types – from familiar standardized tests and classroom quizzes, to coached practice systems and simulation-based assessments, to portfolios and student-tutor interactions.

The key innovations in ECD are clear conceptual models for Student Competencies, Evidence, Tasks, and Knowledge Domains to guide assessment development, implementation, and delivery:

---

A process like ECD would be most helpful in developing more effective assessments and addressing special assessment needs, such as defining clear use cases and determining whether evidence supports using a given assessment for the intended use case.

4. Develop a Plan for Dramatically Increasing the Use of High Quality, Authentic Performance Assessments and Comparable Rubrics

There are very few barriers to effectively implementing high-quality performance assessments. As Stuart Kahl of Measured Progress has written:\(^{222}\):

Effective performance assessment, neglected for some time, is on the rise. Fortunately, with greater awareness of the concept of alignment of tests to standards, current underused capabilities of technology and lessons of the 1990s authentic assessment era, we know how to do it.

What are those performance assessment lessons learned from past experience?\(^{223}\)

Creating reliable, valid, feasible and cost-effective performance assessments can be developed with attention to these topics:

- **Careful task design** based on a clear understanding of the specific knowledge and skills to be assessed – representing important disciplinary content that represents core concepts and abilities validly – and an understanding of how students develop cognitively.
- **Reliable scoring systems** based on standardization of tasks and well-designed scoring rubrics, training of scorers, moderation of the scoring process to ensure consistency in applying the standards, and auditing of the system to double-check and upgrade comparability.
- **Methods for ensuring fairness** based on the use of universal design principles, careful linguistic choices to avoid sources of confusion unrelated to the content being measured, cultural review of items, and pilot testing of tasks to see how they perform with different test takers.
- **Effective use of technology** to deliver and administer assessments; enable simulations, research tasks, and other sophisticated assessment opportunities; adapt assessments to better measure student abilities and growth; and support both human scoring and machine scoring of open-ended items, which is becoming more reliable and effective.
- **Professional development** that enables educators to build, use, and score assessments that will inform and guide their teaching.
- **Administrative support** from education agency officials and legislators at the state and federal levels, offering targeted assistance to teachers, administrators, and school systems that allows their effective participation in new assessment systems and leverages improvements in teaching.
- **Proper use** that reveals areas needing improvement and leads to curriculum and professional learning supports. Tools such as task blueprints, rubric


specifications, and training and scoring protocols should be developed to support the proper use of performance assessments.

These and other recommendations outlined in this report and others’ can be used to create a comprehensive plan for furthering authentic performance-based assessments.

5. Create Systems of Assessment Involving Triangulation of Multiple Assessment Methods and Integrated Visualizations of Learning Progress

Research and develop ways to integrate evidence of learning progress from multiple assessment methods and from multiple assessment instruments measuring related competencies, as well as varieties of assessment results from across all four educational dimensions. A more holistic view of learning progress can be captured visually as in this example of “radar” display of learning progress:

6. Ensure that All Assessments Provide Guidance for Performance and Practice Improvement
For assessments to go beyond sorting to supporting both improved learning and effective teaching, strategies for guiding learner and teacher improvement must be a part of every assessment tool, instrument and program – assessments for learning and improving.

7. Promote the Research and Development of a Wider array of Simulation and Game-based, Technology-empowered “Stealth” Assessments
Create a wider array of simulation- and game-based embedded “stealth” assessments that use a variety of assessment methods, including Situational Judgment Tests, Virtual Reality, etc.

8. Develop Guidance for Effective 21st Century Assessment Implementation
The Advisory Panel developed a rough draft of an implementation guide to begin this process – since assessment implementation is so critical to the effective use of assessments. This initial effort can be deepened and expanded to cover implementations of the full range of assessments for improving learning and transforming education systems, involving all education stakeholders including parents, business and community members and elected officials.

9. Analyze, Adapt and Prioritize OECD’s and other Policy Suggestions for Meeting the 21st Century Assessment Challenge
Based on a 5-year study of assessment practices in 28 countries (OECD Review on Evaluation and Assessment Frameworks for Improving School Outcomes, 2009-13), the OECD’s report of findings (OECD, 2013, Synergies for Better Learning: An International Perspective on Evaluation an Assessment, Executive Summary) offers policy recommendations for 21st century assessment systems. These recommendations, and others referred to in this report, can be further evaluated, adapted, prioritized and strategized for possible policy initiatives to be developed and implemented by the ARC Consortium.

B. FRAMEWORK-RELATED ASSESSMENT RECOMMENDATIONS

1. **Develop an R&D Agenda that Prioritizes the Twelve Elements of the CCR Framework, Identifies which Organizations would be the Best Candidates for the R&D Work, and Includes a Resource Development Strategy for Ongoing Support**

All 12 of the CCR Framework elements will need further research and development to create a solid foundation for the evolution of assessment into the future. ARC Members will be essential in determining the appropriate priorities and sequences for this work that will best serve their needs and the needs of the assessment field, and can help direct, support and oversee this work through their participation in Working Groups.

2. **Create a Living Online Repository of High-Quality Assessments Aligned to the Four-Dimensional Framework**

To create an online repository, a database of categories and criteria similar to those used in this report will need to be created based on best practices in assessment review. It can then be used to search and filter assessments in the repository, with a collaboration component to discuss and comment on the ratings and real-world user experiences with the assessments.

3. **Create a Plan for Technology-enabled Teacher, Trainer and Leadership Development to Align Curriculum, Assessment Implementation, Professional Development, Infrastructure and Policy with Four-Dimensional 21st Century Assessments**

In a later phase of ARC work, it will be important to create a plan for collaborative creation of educational leadership and teacher professional development guidance that will increase the capacity and competencies of educators to shift assessments and teaching/learning practices toward a 21st century approach.

One innovative strategy or teacher development would be to promote the use of educator micro-credentials that provide performance tasks for both teachers and students, with their combined work submitted as evidence of teacher competency, judged by expert educators.

4. **Develop a Public Education, Outreach and Public Relations Campaign to Inform and Promote Four-Dimensional 21st Century Learning and Assessments**

Also in a later phase of ARC work, it will be advantageous to develop a broad public relations plan for increasing awareness, understanding, and support among all education constituents for the needed shifts in assessment and learning practices. Promoting the integration of all four dimensions of learning and assessing will be necessary to gain increasing support for the further research and applications needed to evolve the field of assessment and to help educational system transformations.
IX. CONCLUSION

Educational assessment has arrived at the threshold of a new era.

To effectively measure and support student learning progress in all four dimensions of learning – Knowledge, Skills, Character and Meta-Learning – the current assessment landscape must transform and innovate its way toward new 21st century perspectives and practices. These approaches will likely include such features as:

- **Deeper Competency Models** – stronger research-based models of the essential components and common progressions of learning competence
- **Richer Performance Evidence** – authentic, performance-based demonstrations of capabilities, using diverse methods and media to capture insightful evidence of learning progress
- **Assessments FOR Learning** – all assessment and evaluative efforts including a focus on supporting and motivating deeper and broader learning progress, beyond traditional student sorting
- **Assessments AS Learning** – embedded “stealth” assessments in online learning simulations and games, and evaluative “lifestream” personal data captured by sensors monitoring daily activities as authentic demonstrations of applied learning
- **Integrating Multiple Methods** – more effective triangulation of multiple assessment methods across the four educational dimensions, with deeper research on how these competency elements interact and can affect each other
- **More Effective Assessment Use** – better alignment between assessment uses and education goals, practices, improvement strategies, and education transformation

CCR’s Assessment Research Consortium (ARC) offers a global level playing field and collaborative platform to help pave the new assessment pathways crucial to ensuring learner success in learning, work, civic, and social life in the 21st century.
X. APPENDICES

1. CCR’s Assessment Research Consortium (ARC) Prospectus

Assessment Research Consortium
Transforming Assessments for a New Education Era

Introduction
The Center for Curriculum Redesign (CCR) is an international convening body and research center whose goal is to expand humanity’s potential and improve our collective prosperity by redesigning K-12 education standards for the 21st century. CCR brings together non-governmental organizations, jurisdictions, academic institutions, corporations, and nonprofit organizations, including foundations, to respond to the question, “What should students learn for the 21st century?” and to openly propagate recommendations and frameworks on a worldwide basis.

Education systems across the globe have been well tuned to the demands of the past Industrial Age, and are now struggling to ready students for success in a rapidly transforming, present and future, Innovation Age. The last major changes to curriculum were effected in the late 1800s as a response to the sudden growth in societal and human capital needs. As the 21st century bears little resemblance to the 19th century, education curricula are overdue for a major redesign, emphasizing depth of understanding and versatility, to meet the needs of our global society.

So far, wave after wave of attempted fixes and improvements to traditional schooling have brought little of the promised Golden Age of Learning. Curricula have been adjusted over time, but have never been deeply redesigned to address all dimensions of education: knowledge, skills, character, and meta-learning.

As a non-partisan, globally inclusive, and independent global organization, CCR is strategically partnering with international and national organizations and jurisdictions to build a global consensus on each of the dimensions and elements of a new framework of goals, measures, and practices that will ready all students for learning, careers, citizenship, family and community life in the 21st century.

Framework of Education Goals

Why an Assessment Research Consortium?
With the development of a new framework and wider goals for education, enhanced sets of measures are now needed to track progress toward those goals. In other emerging fields and industries, new standards for measurement, evaluation, and assessment of progress are often established by a pre-competitive, collaborative consortium of organizations and experts that collectively create the “level playing field” of research, high standards, and effective practices that will best serve all constituents.
Presently, there are numerous assessment efforts around the world that are disconnected from each other, and as a result critical mass is not reached and progress is stymied. As in other industries such as semiconductors, biotech and many others, this consortium aims at harmonizing the very many disparate research efforts, and providing a critical mass behind such complex research by sharing the costs and the outputs on a pre-competitive basis. Once the foundational research, standards, and exemplary practices are firmly in place and shared among all consortium members, it is then time to let a “thousand innovative flowers bloom” both collaboratively and competitively in the global “coopetition” market for services and products.

The goal of an education assessment consortium is to establish a collective field for redesigned systems of measuring student, classroom, school, district, regional, state, national, and international progress in learning, aligned to 21st century global goals and desired education outcomes.

How will the Assessment Research Consortium Work?
Leaders from government, the private sector, academia, and nonprofit organizations are invited to join the consortium, which will collectively oversee key research projects to define “assessments for and as learning” across the CCR Framework’s four dimensions of education: knowledge, skills, character, and meta-learning (for more details, see “Research Strategy” document).

### Assessments OF Learning
- Standardized, psychometrically sound tests or tools for measuring whether students have developed knowledge, skills and other competencies compared to established standards, benchmarks and learning goals for the purpose of accountability, program evaluation, or research
- *Example: US NAEP Test*

### Assessments FOR Learning
- Formative and some portfolio summative methods of identifying: student learning progress in ongoing work and performance tasks; new learning needs as they arise; and opportunities to revise work and improve competencies
- *Example: Performance Task*

### Assessments AS Learning
- Mostly formative, meaningful learning tasks with embedded assessments that provide immediate feedback as part of the ongoing learning experience, with a progression of challenges for increasing mastery with a wide variety of feedback
- *Example: Online learning game*

The consortium aims to produce cutting-edge recommendations on assessments related to the framework’s 12 competencies. Given the urgency of aligning education with 21st century societal needs, the consortium aims to complete the assessment recommendations within a 3-5 year period and advocate for rapid adoption.

Members of this consortium will have the opportunity to provide direct input on strategy and serve on project teams that will explore each of the 12 competencies. Project teams will be led...
by a member of the consortium, define goals for the project, identify researchers to undertake
the work, and oversee projects, which will be funded by foundations, governments and other
donors. Members of the consortium will have early access to the research results for up to 2
years, after which the material will be publicly available for nonprofit use across the world.

Supporting the work of the consortium will be a small team of experienced consultants, who will
oversee all research projects (in conjunction with team leads), handle administration, and
support a small Advisory Board. The Advisory Board will be chaired by Charles Fadel, founder
of the CCR, and include representatives from the membership (for more details, see “Operating
Protocols” document).

**Why Join this Consortium?**
The Assessment Research Consortium aims to develop the most relevant education
assessments for the 21st century on an assertive timeline.

*Members of the consortium will benefit from:*
  - Engaging with some of the most respected, entrepreneurial leaders from all sectors who
    are invested in aligning education with society and industry’s current and future needs
  - Contributing their thought leadership to a global pioneering effort to change the culture
    and course of learning assessments
  - Using the “what gets measured, gets taught” power lever to transform education practice
  - Collectively influencing education policy reform at many levels
  - Helping to set a level playing field for a new era of educational measures and
    assessments
  - Having important input on the future of international assessments, such as PISA
  - Gaining early access to research findings, up to 2 years in advance of public release
  - Branding their organization as a leader in developing 21st century assessments

*This consortium is unique in many ways:*
  - *Transparency* – The consortium is designed for maximum transparency of goals,
    research results, and recommendations
  - *Speed of decision-making and research* – Given the urgency of improving educational
    models and assessments, the consortium has assertive goals for accomplishing its
    research and recommendations. The Advisory Board will be small, decision-making will
    be streamlined, and there will be obsessive attention to meeting tight deadlines
  - *Global reach* – The work of the consortium is focused on setting global standards for
    assessment that are not constrained by national or political agendas
  - *Scope* – The consortium’s work is focused on deeply analyzing all aspects of education
    that are essential for student success by examining skills, knowledge, character, and
    meta-learning in a manner not being undertaken by any other institution or government
  - *Relevance* – Research will focus on linking K-12 education to current and future
    workforce development needs
  - *Knowledge* – Connection to knowledge domains, wherever possible
- *Resources* – The consortium will serve as an aggregator of the best thinking, research, and resources on assessments for the next generation of students

**Membership in the Consortium**
Members of the consortium will include representatives from:
- Ministries, Departments of Education and Education Agencies
- Departments of international organizations
- Assessment development, training and research organizations
- University Schools of Education that have deep expertise in measuring and motivating learning
- Workforce development organizations
- Corporate Human Resources and Social Responsibility departments
- Curriculum development and training organizations
- Professional teacher and leadership development organizations
- Education institutions from pre-K schools through University
- Informal learning organizations (after-school, museums, online programs, etc.)
- Education-focused foundations and international development organizations

**Funding of the Consortium**
Membership dues will provide funding for a small team of consultants who will oversee the research projects and administration for the consortium. The lead consultant will be responsible for working closely with CCR to secure funding for the consortium’s research projects. Initial funding for the inception phase of the consortium has already been provided by the Argosy Foundation, the Koshland Family Foundation and the Oak Foundation.

CCR’s work has already been supported by, among others:
- The Bill & Melinda Gates Foundation
- The Hewlett Foundation
- The Nellie Mae Education Foundation
- The Simons Foundation
- The Henri Moser Foundation
- Fondation Helvetica Education
2. Review of Assessment Principles

We often think of assessment, whether of groups or individuals, as divided into two categories, summative and formative; these two categories often, though not always, overlap with so-called high stakes and low stakes testing.

Another way to think of measurement is as being “assessment of learning” -- after the fact, in summation, and usually for accountability, program, or student evaluation, and hence, high stakes; “assessments for learning” -- administered throughout the course of learning, supporting student growth and teacher methodology, frequent and immediate, and so usually formative and low stakes; and “assessments as learning” -- embedded in the course of student learning, non-interruptive, engaging and instructive in itself, and it should be noted, of potential for both of and for learning, summative and formative.

The following, mostly current, resources and guides further illuminate and inform our thinking about the purposes and future of assessment.

*Bryk/Carnegie Foundation (2015)*

In their 2015 book, *Learning to Improve: How America’s Schools Can Get Better at Getting Better*, Bryk et.al suggest we think of group measurement as having three main purposes: measurement for accountability; measurement for research; and what they think is most in need of development and implementation, measurement for improvement.

In contrast to the other two measurements, measurement for improvement entails

- more frequent measurement
- determining whether an educational change is working, in real time or close to it,
- being easily embedded in day to day work
- signalling actionable change
- educators as primary users
- data sharing in low stakes, low risk, safe environment conducive to change.

*Duckworth/Yeager (2015)*

In a 2015 AERA paper, Duckworth and Yeager argue that there are five primary purposes, or uses for, educational measurement, each with their own concerns regarding validity and reliability.

1. Research: Research studies have sought to reject the null hypothesis of no relation between personal qualities and later life outcomes, under testing conditions where incentives to distort responses were minimal-- a very different project than the applied uses below.
2. Program Evaluation: Many educational programs including charter schools, in-school programming, and after school activities, aim to cultivate personal qualities.
3. Accountability.
4. Individual diagnosis: for tracking or remediation decisions, or potentially for selection and honors.
5. Practice Improvement: to systematically improve personal qualities across contexts.

*Gordon Commission (2013)*
The Gordon Commission on the Future of Assessment in Education, (2013), calls for a new generation of assessment practices and tools, and makes the following arguments:

1. Assessment must be transformed to better support teaching, learning, and human development.
2. Assessment must fully represent the competencies that the complex world demands.
3. Assessments must be models worthy of attention and energy of teachers and students.
4. Assessment must be for more than accountability, it must be for providing instructionally relevant feedback to teachers and students.
5. Systems of assessment must provide accountability and feedback in synergistic ways.
6. Assessments designed for one purpose (e.g. accountability) are seldom best suited for other purposes (instructional adjustment).
7. Accountability must be achieved in a way that supports high quality teaching and learning.
8. What we choose to assess is what will end up being the focus of classroom instruction.
9. New assessment resources and tools are needed that better integrate with classroom teaching and learning, and better represent current thinking on how students learn and on changes in the world at large.

Consequential validity is critical to add to the traditional forms of measurement evaluation, reliability and validity (each of which, it should be added, have multiple key sub-categories).

"what has received far less attention is a different principle, the degree to which evidence stimulates improvement and therefore become consequential…. Consequential validity posits that assessment must be valid for the purposes for which it is used, consistent with relevant professional standards, and-- this is the key point-- that the impacts of consequences of its use should be factors in determining validity. This focus on consequences underscores the fact -- known all too well by anyone who has worked for long in assessment-- that the relationship between evidence and action is not always neat, rational, or linear.

"Consequential validity thus brings an important new lens to questions about evidence by putting a premium on the productive use of information critical to student learning outcomes assessment."

* Pearson (2014)
A 2014 Pearson Paper by Hill and Barber “Preparing for a Renaissance in Assessment” argues that assessment is in great need of reform-- and that these reforms are underway and emerging-- and calls for an ideal composed of five elements:

1. Assessments that can accommodate the full range of student abilities. (And not be attentive only to the middle of the the distribution, nor be too narrowly attentive to cut scores).
2. Assessments that provide meaningful information on learning outcomes (and not be overly reliant on levels that reveal little about what a student can do, or provide information too late or too general to be of value for improving learning, or be reduced to just a mean or cut score percentage).
3. Assessments that can accommodate the full range of value outcomes (and not just cognitive/academic achievement narrowly defined and narrowly measured).
4. Assessments that support students and teachers in making use of ongoing feedback to personalize instruction and improve learning and teaching.
5. Assessments that have integrity and that are used in ways to motivate improvement efforts and minimize opportunities for cheating and gaming the system.

For our purposes, this review concerns itself first with ideal number 3—assessing a wider range of outcomes than is the norm, and secondarily with number 2, 4, and 5.

We summarize what is known about test-induced changes in instructional practice. Research identified changes in what is taught, i.e., the curriculum, as a result of assessments. These effects include changes in the order in which content is covered, narrowing the focus of the curriculum to certain subjects or content, and focusing on particular types of skills (e.g., basic skills or facts, higher-order thinking skills or concepts).

Perhaps the most commonly reported reactions to tests involve reallocation of curriculum content to focus more on tested subjects or topics and less on subjects or topics that are not tested. The tendency for educators to focus more on tested than non-tested content could be considered beneficial in the context of a testing program that covers a broad range of skills and knowledge, but would generally be viewed as undesirable if it occurs in response to tests that sample only a subset of the skills and knowledge that are considered important. While reduction in emphasis on social studies, art, and other subjects that are frequently omitted from high-stakes testing programs typically receives the bulk of attention from critics of testing, both forms of narrowing—across and within subjects—have been documented in the literature, and both raise concerns about what students are missing (House of Commons, Children, Schools and Families Committee [United Kingdom], 2008; Yeh, 2005).

There is a large amount of research suggesting that high-stakes testing leads to increased focus on basic skills (Jones et al., 1999; Herman and Golan, 1991; Shepard and Dougherty, 1991) or facts (Johnston and McClune, 2000; Gallagher and Smith, 2000). Watanabe (2007) noted how middle school English teachers in North Carolina tend to frame their questioning in terms of right and wrong answers to parallel the multiple-choice state test.

Testing not only influences what teachers teach, but in some cases it can affect how they teach. Several studies identified changes in the ways teachers convey content in their classrooms. These include engaging in test-preparation activities, adopting new instructional strategies, and changing assessment practices.

Specifically, the literature suggests that new, CCSS-aligned assessments are likely to promote desirable changes in practice when the following conditions are met.

1. Test content and format should mirror high-quality instruction
2. Tests should be used only for purposes for which they were designed and validated.
3. Score reporting should be optimized to foster instructional improvement.
4. Teachers should receive training and support to interpret and use test scores effectively.
5. The test scores should “matter,” but important consequences should not follow directly from test scores alone.
6. If there are externally mandated, high-stakes tests, they should be part of an integrated assessment system that includes formative and summative components.
7. Accountability metrics should value growth in achievement, not just status, and should be sensitive to change at all levels of student performance, not just changes at a single cut point.

8. Assessment should be one component of a broader systemic reform effort.

* Wiggins and McTighe’s *Schooling by Design* (2007)

"First and foremost, academic leaders need to ensure that every educator understands that his or her job is to work toward the mission and goals by identifying and working to close the inevitable gaps between mission and reality, between desired learning results and actual performances by students on measures that matter."

* Balanced Assessment System:

The term "balanced assessment" refers to the strategic use of formative, interim, and summative measures of student performance in a way that addresses immediate student needs, informs ongoing instructional adjustments, and guides long-term educational improvement. A balanced approach to assessment recognizes both the strengths and limitations of the various assessment types and highlights the most relevant assessment data to address student learning needs, improve instruction, and increase accountability at all levels within an educational system.

A balanced assessment system is a synthesis of interacting quality assessments and assessment methods that inform instruction, enhance student learning and engagement, and guide continuous educational improvement. These assessments align to rigorous and relevant learning outcomes, provide essential student performance data that is both valid and reliable, and support all stakeholders in taking ownership of monitoring student progress and ensuring academic success.

While the philosophy behind Balanced Assessment certainly is not new to education, the call for a system that uses multiple and varied measures of student performance has grown louder in recent years. While the push for higher levels of academic achievement and accountability continues to increase, more people have realized that a single test cannot provide a comprehensive evaluation of student performance. The state of Colorado has even gone as far as passing legislation that requires the inclusion of multiple student performance measures in teacher evaluations as well as the Unified Improvement Planning process for both schools and districts.

* Interpretation Use Argument Approach to Validity (Kane, 2013)

To validate an interpretation or use of test scores is to evaluate the plausibility of the claims based on the scores. An argument-based approach to validation suggests that the claims based on the test scores be outlined as an argument that specifies the inferences and supporting assumptions needed to get from test responses to score-based interpretations and uses. Validation then can be thought of as an evaluation of the coherence and completeness of this interpretation/use argument and of the plausibility of its inferences and assumptions.

1. It is the proposed score interpretations and uses that are validated and not the test or the test scores.

225 https://sites.google.com/a/dcsdk12.org/bas/
2. The validity of a proposed interpretation or use depends on how well the evidence supports the claims being made. 
3. More-ambitious claims require more support than less-ambitious claims. 
4. More-ambitious claims (e.g., construct interpretations) tend to be more useful than less-ambitious claims, but they are also harder to validate. 
5. Interpretations and uses can change over time in response to new needs and new understandings leading to changes in the evidence needed for validation. 
6. The evaluation of score uses requires an evaluation of the consequences of the proposed uses; negative consequences can render a score use unacceptable. 
7. The rejection of a score use does not necessarily invalidate a prior, underlying score interpretation. 
8. The validation of the score interpretation on which a score use is based does not validate the score use.
### 3. Selected Assessment Compendia

<table>
<thead>
<tr>
<th>Report Name</th>
<th>Web Address</th>
<th>Source</th>
<th>Measures</th>
<th># of Measures</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Sedlacek Surveys</strong></td>
<td><a href="http://williamsedlacek.info/publications.html#Surveys/Instruments">http://williamsedlacek.info/publications.html#Surveys/Instruments</a></td>
<td>U. of Maryland Baltimore researcher W. Sedlacek, author of Beyond the Big Test, Noncognitive Assessment in College and University</td>
<td>Positive Self-Concept, Realistic Self-Appraisal, Preference for Long Term Goals, Leadership</td>
<td>One main one, plus many others</td>
</tr>
</tbody>
</table>
4. Technical Analysis Criteria Guide

Technical Criteria

<table>
<thead>
<tr>
<th>Content Validity</th>
<th>Predictive Validity</th>
<th>What is Predicted? (Fill-in)</th>
<th>Internal Consistency</th>
<th>Repeatability</th>
<th>Type of Repeatability (Test-Retest, Interrater, Parallel Forms)</th>
<th>Fairness/Equity</th>
</tr>
</thead>
</table>

**Content Validity**

*Parameter to look for*

There is no quantitative parameter for content validity; it is purely a qualitative judgment.

*Proposed Scale*

High: Assessment Items are explicitly cross-referenced to existing state, national, or international standards or have been positively reviewed by a panel of third-party subject-matter experts with established credentials in the field.

Medium: An external advisory panel with some training or experience in the field has concluded, based on reading typical items, that they reflect the content purported to be covered by the advertised assessment topic, and that the items are not posed in a "leading" or biased way.

Low: An external advisory panel with some training or experience in the field has concluded, based on reading typical items, that some/many items are not germane to the advertised topic, or that some/many items are posed in a leading or biased way.

None: A credible panel of any type has concluded that the items in the assessment bear no apparent relationship to what the assessment claims to measure. An example would be if a test that claimed to measure reading skills contained only mathematics problems. Such a situation is very rare and indicates that most likely, the assessment was just miscategorized.

**Predictive Validity**

*Parameter to look for*

$r$, the correlation coefficient between the assessment score and an external measure of interest (e.g., wages, years of education, personal happiness on a 1-7 scale, etc.)

*Proposed scale*

High: One or more research studies, representing a combined total at least 300 human subjects, has shown the assessment correlates with an external outcome of interest
(e.g., employment, wages, graduation rates, incarceration rates) with an $r > 0.35$ (or, for a negative outcome, $r < -0.35$).

Medium: One or more research studies, representing a combined total at least 40 human subjects, has shown the assessment correlates with an external outcome of interest (e.g., employment, wages, graduation rates, incarceration rates) with an $0.2 < r < 0.35$ (or, for a negative outcome, $-0.35 < r < -0.2$). This correlation coefficient is also called the “criterion validity.”

Low: Research studies show the correlation coefficient between assessment score and outcome of interest is $r < 0.2$ (or, for a negative outcome, $r > -0.2$). Also, this category covers situations where no or few reported data exist showing a correlation between the assessment results and an external outcome of interest.

None: The assessment’s makers have issued strong claims of predictive validity, but independent studies have debunked the claims. The correlation between assessment scores and the claimed external outcomes is either zero ($-0.05 < r < 0.05$), or statistically indefensible ($p > 0.3$; i.e. more than 30% likely to have occurred by chance), or in the opposite direction of what was claimed.

Asterisk: An asterisk is used if some or all of the data required to grant a high or medium rating are supplied by the test manufacturer.

Technical note: The correlation coefficient between assessment scores and an external outcome of interest is sometimes called “criterion validity.” The breakpoints for criterion validity were chosen based on the US Dept. of Labor’s *Testing and Assessment: An Employer’s Guide to Good Practices* [1], which summarizes the “practical” state of the art in cognitive and non-cognitive testing.

*What is Predicted?*

This pertains to predictive validity and is filled in only if there is a Predictive Validity rating.

**Internal Consistency**

*Parameter to look for*

Cronbach’s $\alpha$ (alpha), the “reliability coefficient.”

*Proposed scale*

High: Cronbach’s $\alpha > 0.8$.

Medium: $0.7 < $Cronbach’s $\alpha < 0.8$.

Poor: $0.5 < $Cronbach’s $\alpha < 0.7$.

None: $Cronbach’s \alpha < 0.5$
Asterisk: An asterisk is used if some or all of the data required to grant a high or medium rating are supplied by the test manufacturer.

Technical note: Cronbach’s $\alpha$ measures the internal consistency of the items in an assessment: do the questions or rated activities hang together as a coherent set of measures? A counterexample would be an exam that mixed questions on, say, English history and accounting, and then claimed it measured just English history. The cutoff points for Cronbach’s $\alpha$ were taken from George and Mallery [2], using their “good” ($\alpha=0.8$) as the boundary between “high” and medium,” and their “acceptable” ($\alpha=0.7$) as the boundary between “medium” and “low.” George and Mallery’s [2] “unacceptable” ($\alpha<0.5$) forms the “none” category. If an assessment intentionally covers multiple concepts, delivering separate subscores for each, then Cronbach’s $\alpha$ should be calculated separately for each of the subsections.

Repeatability (including Test-Retest Reliability, Parallel Form Reliability, and Interrater Reliability)

Parameter to look for

- $r$, the correlation coefficient between the first and second measurement on the same population.

Proposed scale

**High:**
- Test-Retest Reliability: $r>0.85$.
- Parallel Forms Reliability: $r>0.85$.
- Interrater Reliability: $r>0.7$ or $\kappa>0.6$

**Medium:**
- Test-Retest Reliability: $0.7<r<0.85$.
- Parallel Forms Reliability: $0.7<r<0.85$.
- Interrater Reliability: $0.5<r<0.7$ or $0.4<\kappa<0.6$

**Low:**
- Test-Retest Reliability: $0.5<r<0.7$.
- Parallel Forms Reliability: $0.5<r<0.7$.
- Interrater Reliability: $0.25<r<0.5$ or $0.2<\kappa<0.4$

**None:**
- Test-Retest Reliability: $r<0.5$.
- Parallel Forms Reliability: $r<0.5$.
- Interrater Reliability: $r<0.25$ or $\kappa<0.2$

Asterisk: An asterisk is used if some or all of the data required to grant a high or medium rating are supplied by the test manufacturer.

Technical note 1: **Test-retest reliability** is a measure of the correlation between test scores when the same exact test is given to the same individuals twice in a row. When different versions of the same test are given to the same individual twice in a row (Form A and then Form B), the correlation between these two scores is called **parallel forms reliability**. Finally, when the repeat measure is two or more raters judging the same individual through personal
observation (e.g., giving 1-10 ratings of ice skating skill using a grading rubric), the correlation between scores is called **interrater reliability**. Interrater reliability tends to be very low compared to the other forms of reliability, which is why its cutoffs are adjusted downwards. Cohen’s kappa (κ) is an interrater reliability measure that will account for the possibility that two raters might agree by chance. Intraclass correlation coefficient (ICC) is another interrater reliability measure but one not often found in assessment documentation.

Technical note 2: All of the rating breakpoints assume that no actual change in a person’s ability, traits, or skills has occurred between the initial measure and the repeat measure. For test-retest measures, the time between repeated testing needs to be sufficient to allow the candidate to forget any answers he may have previously memorized, but not so great that changes in the underlying construct may have occurred.

Technical note 3: The value of r=0.7 is often cited as the benchmark for “acceptable” test-retest and alternate form reliability based on rules of thumb given by industrial-organizational psychology textbooks [3]. This becomes the anchor of the proposed rubric categorization for reliability, occupying the floor of the “medium” category. By dividing the range of r values which are “acceptable” or higher in two, a “high” category (0.85<r<1) is established distinct from the “medium” category (0.7<r<0.85). The r=0.5 cutoff for “none” was chosen to be consistent with common usage for “unacceptable” reliability (e.g., see Ref [4]). In-between “none” and “acceptable” is a grey zone of questionable reliability (0.5<r<0.7), occupying the “low” category in the current rubric.

Technical note 4: The high-medium-low breakpoints for Cohen’s Kappa were informed by Landis and Koch [5], as reviewed by Hallgren [6]. Landis and Koch maintain that 0.0 to 0.2 indicates slight agreement, 0.21 to 0.40 indicates fair agreement, 0.41 to 0.60 indicates moderate agreement, 0.61 to 0.80 indicates substantial agreement, and 0.81 to 1.0 indicates near-perfect perfect agreement. The r values for interrater reliability were adjusted upwards of these to reflect the fact that r is a more generous measure.

**Fairness/Equity**

*Parameter to look for*

Cohen’s d, the effect size.

*Proposed scale*

**High:** d<0.25. Over 90% overlap between protected and non-protected populations.

**Medium:** 0.25<d<0.5. 80-90% overlap between protected and non-protected populations OR alternatively, the assessment tool has been reviewed, edited and approved by an outside panel of gender/racial/ethnic diversity experts.

**Poor:** 0.5<d<3.3. 10-80% overlap between the protected and non-protected populations.

**None:** d>3.3. <10% overlap between the protected and non-protected populations.
Asterisk: An asterisk is used if some or all of the data required to grant a high or medium rating are supplied by the test manufacturer.

Technical note: Cohen’s d measures how far apart two populations are, in units of standard deviation. It is the difference between the two populations’ means, divided by their pooled standard deviation (if the two standard deviations are close, as they often are, then an average standard deviation can be used for the denominator). Thus, a Cohen’s d of 0.5 says that two populations are half a standard deviation apart. The question then becomes, how far apart can the means in assessment scores be, before two populations end up being treated differently? Is it OK for blacks and whites’ test scores to be 1 standard deviation apart? 0.2 standard deviations apart? The EEOC has wrestled with this question in the context of employment testing and decided on an 80:20 rule [7]: any assessment used for employment purposes has to result in protected minorities achieving at least 80% of the success rate as whites. In practical terms, this condition should always be satisfied if the groups’ test score distributions overlap by 80%. To convert an 80% overlap to a Cohen’s d, we use the 80% curve on the graph found at Reference [8] and assume that the two populations have similar standard deviations (SD2/SD1=1, a condition that is close to being observed practice). This calculation says that a Cohen’s d of 0.5 would result in 80% overlap between two populations. We therefore set d=0.5 as the lower boundary of our acceptable “medium” range, with lower values of d corresponding to “less-than-acceptable” instruments. The breakpoint for “high” value was arbitrarily set as the midpoint between a d of 0.5 and a d of 0 (perfect overlap in the two populations’ test scores). This corresponds to a 90% overlap between two normal distributions with equal standard deviations. The breakpoint for “none” was set to correspond to a Cohen’s d of 3.3, i.e., <10% overlap between two populations, as per Ref [8].

It should be noted that real differences do exist between ethnic and racial groups, especially for constructs like literacy, that are impacted by poverty. In this case, it becomes impossible for the test instrument itself to be race-neutral. Indeed, if the instrument is being used for diagnostic and formative purposes, it will be important for the instrument to pick up the differences that do exist, so that an appropriate remediation path can be developed. In these situations an alternative approach can be adopted, in which an external panel of diversity experts, representing the range of populations to be assessed, reviews every item in the assessment for bias and adjusts the instrument until, in their opinion, it is racially and ethnically neutral.

References


