GELP New Metrics working group

This is a short collection of new approaches to measurement and metrics, compiled to exemplify the range of tools and system strategies that are currently in use or becoming available. It is intended to allow us to have richer conversations about new metrics.

The final part offers a frameworks for considering how to bring an evaluative lens to this kind of collection: what kind of assessments for suitable for a high-stakes purpose like end of school qualifications? What composition of assessments should be part of a next generation system of assessment, certification, and reporting?

Part 1: thinking differently about summative assessment

Performance assessments, systems of assessment, and blurring the formative/summative divide

Intro: competencies and performance assessments

In general conversation, the term 'competencies' is sometimes used to refer to broad, overarching capacities such as problem-solving or collaboration. For the requirements of policy documents, however, the term 'competency' typically refers to "explicit, measurable, transferable learning objectives". As such, a competency is usually a subject-bounded combination of skills and knowledge. For example, in Maths one key competency might be *'Students will demonstrate the ability to create and use algebraic models to connect mathematical concepts and properties when solving real-world problems.'*

The difference between a competency and the kind of learning outcome or standard found in traditional curricula, is that a competency emphasizes both the application or creation of knowledge, and tends to represent an amalgam of 'know-that' and 'know-how', as opposed to being a discrete 'fact' or 'skill'.

What is the relationship between competencies and performance assessment?

Because competencies are a composite of knowledge and skills, they are generally assessed using what is often referred to as a performance assessment. Performance assessments can be more or less authentic/real-world – for example, some are carried out under test conditions, and some take the form of a piece of school work; others take the form of a 'final product' or are presentations of learning in front of a real audience. Typically, any kind of performance assessment comes with a rubric or similar criteria for assessing the quality of the product or performance.

1.1. New Hampshire Competency frameworks

In 2005, New Hampshire became the first U.S. state to ordain that high school graduation should be based on demonstration of competency, as opposed to 'seat time'. Each district was given responsibility for developing 'competency pathways' by which students would demonstrate their mastery of relevant course content and skills and proceed to graduation. Over time, a central process has developed for reviewing and accrediting competency pathways, to try and improve quality across the state.

New Hampshire 'competency validation rubric' – a set of criteria for establishing the strength or weakness of proposed competency statements:
http://www.education.nh.gov/innovations/hs_redesign/documents/validation_rubric_for_course-level-competencies.pdf

2013 frameworks for Maths and English Language competencies developed in New Hampshire:
http://www.education.nh.gov/competencies/index.htm

More details:

A simple explanation of competencies, as they are typically understood in the U.S. context:
http://www.competencyworks.org/2013/04/the-three-legged-stool-of-competency-frameworks/

Detail of competency-based approaches as developed in New Hampshire and gradually in other U.S. states:
https://sites.google.com/site/competencybasedpathways/home/understanding-competency-based-approaches

Key papers on developing competency frameworks:
https://sites.google.com/site/competencybasedpathways/key-issues/developing-competencies

1.2 Performance assessment in Colorado

The Colorado department of Education promotes performance assessment as the summative assessment method for ascertaining competency throughout school education. The Colorado Professional Learning Community, an online space created through collaboration between the DOE and several professional associations, has compiled resources for educators to described and support the development process for performance assessments:

http://www.coloradoplc.org/node/12765

The definition of Performance Assessment used in Colorado -

An assessment based on observation and judgment. It has two parts: the task and the criteria for judging quality. Students complete a task (give a demonstration or create a product) and it is evaluated by judging the level of quality using a rubric. Examples of demonstrations include playing a musical instrument, carrying out the steps in a scientific experiment, speaking a foreign language, reading aloud with fluency, repairing an engine, or working productively in a group. Examples of products can include writing an essay, producing a work of art, writing a lab report, etc.

(Pearson Training Institute, 2011)

Specific resources:

An overview of performance assessment tasks:
http://www.coloradoplc.org/files/archives/a._performance_task_development_presentation_0.pdf

On rubric development:
http://www.coloradoplc.org/files/archives/a._rubric_development_presentation.pdf

1.3 NY Performance Standards Consortium

This consortium of small public and alternative high schools formed in 1997 to develop alternatives to the New York state 'Regent's Exams' – high-stakes exams required of students to graduate.

Consortium website: http://performanceassessment.org/

The consortium schools each have a waiver allowing them to graduate their students based on demonstration of mastery through performance-based assessments – that is, assessments that take the form of holistic tasks such as writing an essay, performing a science experiment, or applying maths skills to solve a problem. The schools have carried out moderation processes to ensure reliability across their methods of assessing graduation-level work. This is managed through the 'Center for Inquiry in Teaching and Learning', the professional learning body supported by the schools. Descriptions of the moderation process and sample student work submitted for the process can be found here: http://performanceassessment.org/performance/pstudentpapers.html

A collection of sample 'interim assessments' (assessment taken during the school year to assess readiness to demonstrate final learning standards) can be found here:
http://performanceassessment.org/performance/psampleinterim.html

The consortium schools campaign for alternatives to external, one-off high-stakes testing, but are also a body of educator expertise in the implementation of performance assessments. The consortium focusses not just on the development of assessment tasks, but advises that quality performance assessment must be embedded in a particular schools culture and system. Their seven components of successful performance assessment for learning are active learning; formative and summative documentation; strategies for corrective action; multiple ways for students to express and exhibit learning; graduation level performance-based tasks aligned with Learning Standards; external evaluators of student work; a focus on professional development.

Full details of the components can be found here:
http://performanceassessment.org/performance/pcomponents.html

1.4 Performance-based assessment system, Lindsay, CA

In 2010 Lindsay school district in California began adopting a district-wide Performance-based assessment system, where students progress through school activities at different rates, and progression is determined by performance in assessments. (This is equivalent to what is in many places referred to as a competency-based or mastery-based system).

Subjects are divided into 'measurement topics' and students progress through topics based on and end-of-topic exams. Students can track their progress through a topic using the 'capacity matrix' which record learning targets and their evidence of meeting those targets (currently a paper-and-pencil tool). The system has so far only been implemented in elementary and middle schools, but is currently being developed with the help of a $10 million federal Race-to-the-Top grant, including building the progression information into an online system.

District website: http://www.lindsay.k12.ca.us/District/Department/689-Performance-based-System

More detail:

http://www.fresnobee.com/2012/12/11/3097706/lindsay-unified-wins-race-to-the.html

1.5 Computer-based systems of assessment

Computer-based systems of assessment do not necessarily entail performance assessment or assessment of competencies, but they may be a fundamental piece of the architecture that enables this kind of assessment at scale. These are two examples of the development of these kind of systems.

The Amplify Early Literacy Assessments are an example of bringing together formative assessment and the potential to aggregate for system level data.

This tool is designed to provide real-time assessments of the reading skill levels of students in kindergarten through third grade.

http://www.cde.state.co.us/coloradoliteracy/readact/assessmenttool

At the other end of the age scale, the International Baccalaureate Organisation continues to work on how to make all of their diploma assessment computer-based, on a global scale. This presentation offers insight into some of the steps involved in this kind of process, and some of the opportunities, in this case in relation to students of visual art.

https://www.ibo.org/ibla/conference/documents/ElectronicassessmentThenext10years.pdf

Part 2: New foci for assessment

Assessing a wider range of competencies: cross-cutting cognitive, interpersonal and intrapersonal capacities

Source details:
Literature Review of non-cognitive skills

[source for 2.1, 2.2, 2.3]

This review, carried out by the London Institute of Education on behalf of the Education Endowment Foundation and the U.K. Cabinet Office, sets out the existing evidence linking an array of 8 key 'non-cognitive skills' to outcomes for young people. In doing so, it reviews the existing measures for each of these skills, and judges the validity and reliability of these measures with a view to evaluating the resulting evidence. The 8 skills examined are:

1. Self-Perceptions
2. Motivation
3. Perseverance
4. Self-Control
5. Metacognitive Strategies
6. Social Competencies
7. Resilience and Coping
8. Creativity

http://educationendowmentfoundation.org.uk/uploads/pdf/Non-cognitive_skills_literature_review.pdf

The examples of measures on the following few slides are selected from this literature review, based on relevance and reliability. Other measures detailed in the paper may be of interested beyond the three listed below, however, like two of these the majority are self report surveys which may not be applicable to high-stakes processes. Very few of these measures have so far been used consistently in school environments.

Source details:
Social and Emotional Learning measures

This paper from the Raikes Foundation outlines ten tools which met quality assurance standards as measures of social and emotional development. Like the paper above, this review was intended to bring together tools to recommend for program evaluation purposes, in this case in the U.S. middle school context.

The tools featured include surveys that can be completed by both learners and educators, and behavioral rating tools.

The tools altogether cover the five key interrelated social and emotional competencies as described by CASEL (the Collaborative for Academic, Social and Emotional Learning). These are self-awareness, self-management, social awareness, relationship skills, responsible decision-making.

Unfortunately, but not unexpectedly given the complexity of these competency areas, each of the tools carries limitations in terms of implementation time, limited focus, or reliability.

Samples of the measures are included in the paper.

http://raikesfoundation.org/Documents/SELTools.pdf

Source details:
Measuring C21st Competencies

[source for 2.5, 2.8, 2.9]

This report from the RAND group, shared in pre-publication form at the GELP New Delhi event, gathers together a range of examples of methods for assessing '21st century competencies', including some of those detailed above, in 'cognitive' 'intrapersonal' and 'interpersonal' competencies.

http://asiasociety.org/files/gcen-measuring21cskills.pdf

Potentially the most important part of the paper, however, is towards the end (p. 38): 'Key Takeaways from an Investigation of Available Measures of 21st Century Competencies'. This includes the following warning: *'Educators (and learning scientists) do not know as much about teaching and learning 21st century competencies as they do about teaching traditional academic content, so expectations for improvement need to be realistic.'*

However it also includes the statement that: *'Acquiring information about students' understanding of 21st century competencies can make educators and students more intentional about improving the competencies'*

- A question to consider is: how can we ensure that assessment – even in high stakes environments – is supportive of the improvement of teaching and learning rather than a technological process applied externally?

2.1 Motivated Strategies for Learning Questionnaire

This is an example of a tool used to measure the intrapersonal capacity for self-efficacy. Most tools applied to these kind of capacities are somewhat contested or controversial – the capacities themselves are psychological constructs which are not always stable. However, to the extent that we can identify such constructs, this tool has proven reliability and validity[1]. Self-efficacy is of interest and it is often seen as a key outcome of schooling (e.g. as a proxy for autonomy), and is associated with good life outcomes. It could also be seen as an important 'process measure' as self-efficacy is a predictor of and mechanism for academic achievement. The MSLG is also of interest as a measure of metacognitive strategies.

(The description below is adapted from the EEF Literature review of non-cognitive skills):

The MSLQ is an 81-item, self-report instrument consisting of 6 motivation subscales and 9 learning strategies scales. The motivation scales tap into three broad areas: (1) intrinsic and extrinsic goal orientation, task value, (2) expectancy (control beliefs about learning, self-efficacy); and (3) affect (test anxiety).

The MSLQ has proven to be a useful tool that can be adapted for a number of different purposes for researchers, instructors, and students[2]. The MSLQ has been translated into multiple languages and has been used by hundreds of researchers and instructors throughout the world.

For a shorter measure with acceptable reliability, the Students' Approaches to Learning (SAL) instrument was evaluated among approximately 4,000 15-year-olds from each of 25 countries (Marsh et al., 2006). The instrument examines 14 different factors, one of which is perceived self-efficacy. The short scale includes only four items asking students about their confidence in their ability to do well on academic tasks.

A comprehensive history and review of the questionnaire can be found here, including the item list: http://www.sp.uconn.edu/~aja05001/comps/documents/MSLQ_Artino.pdf

1. Pintrich, P. R., Smith, D. A. F., Garcia, T., & McKeachie, W. J. (1993). Reliability and predictive validity of the Motivated Strategies for Learning Questionnaire (MSLQ). Educational and Psychological Measurement, 53, 801-813.

2. Duncan, T. G., & McKeachie, W. J. (2005). The making of the motivated strategies for learning questionnaire. Educational Psychologist, 40, 117-128.

2.2 The Grit Scale

Grit, or perseverance[1], is of interest to educators as a key predictor of life success, with particular relevance to goals such as adaptability and overcoming social disadvantage. The grit scale was developed by Angela Duckworth and colleagues to explore the relative impact of 'grit' on academic life outcomes.

The original scale was self-report scale (Grit-O) was a self-report scale with twelve items. It consisted of two dimensions: Interest and Perseverance of Effort. Responses together gave a measure of 'stamina' in each of these dimensions. In 2009, Duckworth created a shortened version (Grit-S) which retains the original two-factor structure with only eight items, while maintaining acceptable internal reliability.

Both scales have been applied in several research studies to ascertain that this construct of grit is a predictor of both academic and positive life outcomes. Varying degrees of effect size are found with different populations. Crucially, it must be noted that the scales were created to assess individual (between person) differences, as opposed to changes in a person's 'grit' over time. They have not yet been

All of the different scales developed by Duckworth and colleagues can be found here: http://sites.sas.upenn.edu/duckworth/pages/research

1. In psychological terms, grit is one aspect of perseverance – it is typically seen as the most person-specific dimension as opposed to more context- or task-variable criteria such as engagement.

2.3 Torrance Tests of Creative Thinking

Developed in the second half of the C20th by Ellis Paul Torrance and associates, these are a series of both pictoral and word-based tasks which give an indicator of creative thinking. Like IQ tests, there are some serious critiques of the extent to which the tests are context-dependent, but they are seen as the most robust existing measure of creative thinking.

Explanation and links to scoring manual:

http://www.indiana.edu/~bobweb/r546/modules/creativity/creativity_tests/torrance_test.html

Alternative methods of assessing creative thinking focus on rubrics for looking at student work. A 2013 book drawing on the work of education psychologists such as David Perkins details this method, and offers a sample 'Rubric for creativity'.

http://www.ascd.org/publications/educational-leadership/feb13/vol70/num05/Assessing-Creativity.aspx

Another longer rubric to apply to creative works can be found here:

http://grantwiggins.files.wordpress.com/2012/02/creative.pdf

2.5 Mission Skills Assessment (MSA)

This set of measures was developed by the Education Testing Service and the Independent School Data Exchange in the United States, in collaboration with a collection of specific Independent schools. The purpose of the measures is to provide some external evaluation and validation for the schools as they seek to fulfil their 'mission statements' – which typically speak more to the holistic development of their students than to narrow achievement on academic tests. The measures therefore seek to establish the extent to which schools are helping students develop:

- Teamwork
- Creativity
- Ethics
- Resilience
- Curiosity
- Time management

The measures have currently been developed only for use with Middle School students, but high school version are in process. The assessment as its strengths and weakness are described in detail in the RAND paper featured above.

Description and links to a detailed presentation: http://indexgroups.org/msa/

Official website (portal not accessible to non-users; access pending) https://missionskillsassessment.org/

2.6 Critical thinking rubrics

Moving to more purely 'cognitive' skills, there are a great many rubrics for assessing critical thinking:

http://www.uni.edu/adp/documents/LinksforCriticalThinkingRubrics.pdf

Many of these have been developed for use with college-age students, some as part of the move by a collaborative of U.S. colleges. The rubrics assess different dimensions of critical thinking and place different weight on associate skills – for example, some are more taxing than others on reading skills.

There is as yet no consensus on which is most robust for the widest possible range of students, however, one example of a popular rubric for assessing critical thinking is one developed by Peter Facione and colleagues, the 'Holistic Critical Thinking Scoring Rubric': http://web.calstatela.edu/academic/aa/assessment/assessment_tools_resources/rubrics/scoringrubric.pdf

2.7 DiscoTests of Complex thinking

Most researchers and educators would agree that the most promising methods for assessing 'higher order thinking skills' such as complex thinking are thus far limited to the application of rubrics, usually to written work where the complexity of thinking displayed can be evaluated.

This kind of evaluation, while it can be useful in developing educators skills and judgment, is expensive to manage at scale if there is a end for external validation. Therefore, computer-based methods are in development in many research groups.

One of the most advanced is the 'DiscoTest', created by the Developmental Testing Service, a group of researchers attached to Harvard University's School of Education. This 'lexical test' is a computer-based assessment where students provide a short written answer to a problem or question. The test is built to recognise certain patterns of thinking in the response and so can assess the 'level of complexity' of thinking in any given answer. By level of complexity is meant things like – what perspectives have been incorporated? What understanding of the construction of knowledge is displayed here? What dynamic aspects to the phenomena are understood?

DiscoTests are intended as formative assessments for use by teachers to establish a students understanding of an area. They can be built around any sort of content, from physics problems to ethical dilemmas. The theory underpinning the test development is that all facts or knowledge operate at different levels of complexity, and that a key part of education is promoting the ability to handle more complex levels of understanding.

More details can be found in this description of the 'Anatomy of a DiscoTest':
https://discotest.lectica.org/visitors/anatomy.php

2.8 Queensland Performance Assessments

[Source: RAND]

The Queensland Performance Assessments (QPAs) are an externally moderated school-based assessment system, which has been developed since the 1970s in response to Australia's high-stakes university entrance exams.

As well as academic knowledge, the assessments seek to measure problem solving, communication, and learning how to learn. The assessment system is built on a purposeful, systematic, and ongoing collection of data on student learning, rather than information from a single point in time as in an examination. The system is designed to create a tighter link between teaching and learning, and testing. All tests—even those used for high-stakes decisions—are developed by teachers, using national standards and with support from psychometric experts at the Queensland Studies Authority (QSA). Teachers also meet across schools in an attempt to ensure that standards are consistent. The approach intends to promotes teacher professionalism as well as aiming to being a more accurate and fair way of assessing students.

The QPAs are loosely defined in terms of format, but tightly defined on scoring. Teachers can develop a test in any format they want, so long as the standards used to determine proficiency are not only clear but also comparable across schools. In a given example, one school could use a multiple-choice history test, while a neighbour school uses an essay and oral presentation to measure the same standard, so long as the teachers (and the QSA) agreed on how proficiency was determined in each case.

Typically, however, most schools tend towards similar forms of assessment, and many draw on provided item banks in developing their tests. To ensure comparability between schools, assessment design always begins with QSA-developed standards and curricula, and involves consistent back and forth between schools and review boards, which are made up of teachers from across the province, to negotiate a consistent set of student classifications for a given test.

2.9 Project Work, Singapore

In 2006, the Singapore A-levels were reformed. The A-levels are tests required of all pre-university students, and are arranged by subject, with different levels (now known as Higher 1, 2 and 3 categories). A set of complete A level qualifications (typically three H2s and one H1) now also requires completion of a group project, along with the other requirements of a 'Knowledge and Inquiry' and Mother Tongue assessment. The results from A-levels, including the project component, are the major determinant of university admission within the country, so are a high-stakes qualification.

The project work is meant to complement these other A level requirements, by measuring *application* of core academic content, along with communication, collaboration, and learning to learn. This final competency is conceptualised as involved learning independently, reflecting on learning, and taking appropriate action to improve.

To carry out the project work, students are placed into groups by their teacher then are free to select a topic of their choosing. A given set of example of past topics include natural forces, momentum, tradition, groundbreaking individuals, and entertainment. Once students select a project, they work for several weeks preparing for the three associated requirements: a written report, an oral presentation, and a group project file. The project file is Singapore's approach to assessing the student's skills in learning how to learn. Essentially, the file represents the student's way to track progress over time and to reflect on challenges and successes. In particular, students analyze three specific artefacts of their choosing from the project, which are chosen to elucidate the thinking behind the project's design. For example, one might choose an early document outlining the argument that will underlie the paper. From there, the student could discuss how the ideas were formulated, or how they evolved over the course of work.

Scoring of each of the three project requirements is conducted entirely by local teachers. Each requirement receives a specific weight in the final grade. Each student in a group receives the same score for the paper and presentation, and an individual score for their project file.

More details – description of the assessment framework and intentions for Project Work, including each component: http://www.seab.gov.sg/aLevel/2013Syllabus/8809_2013.pdf

Part 3: from tools to a system

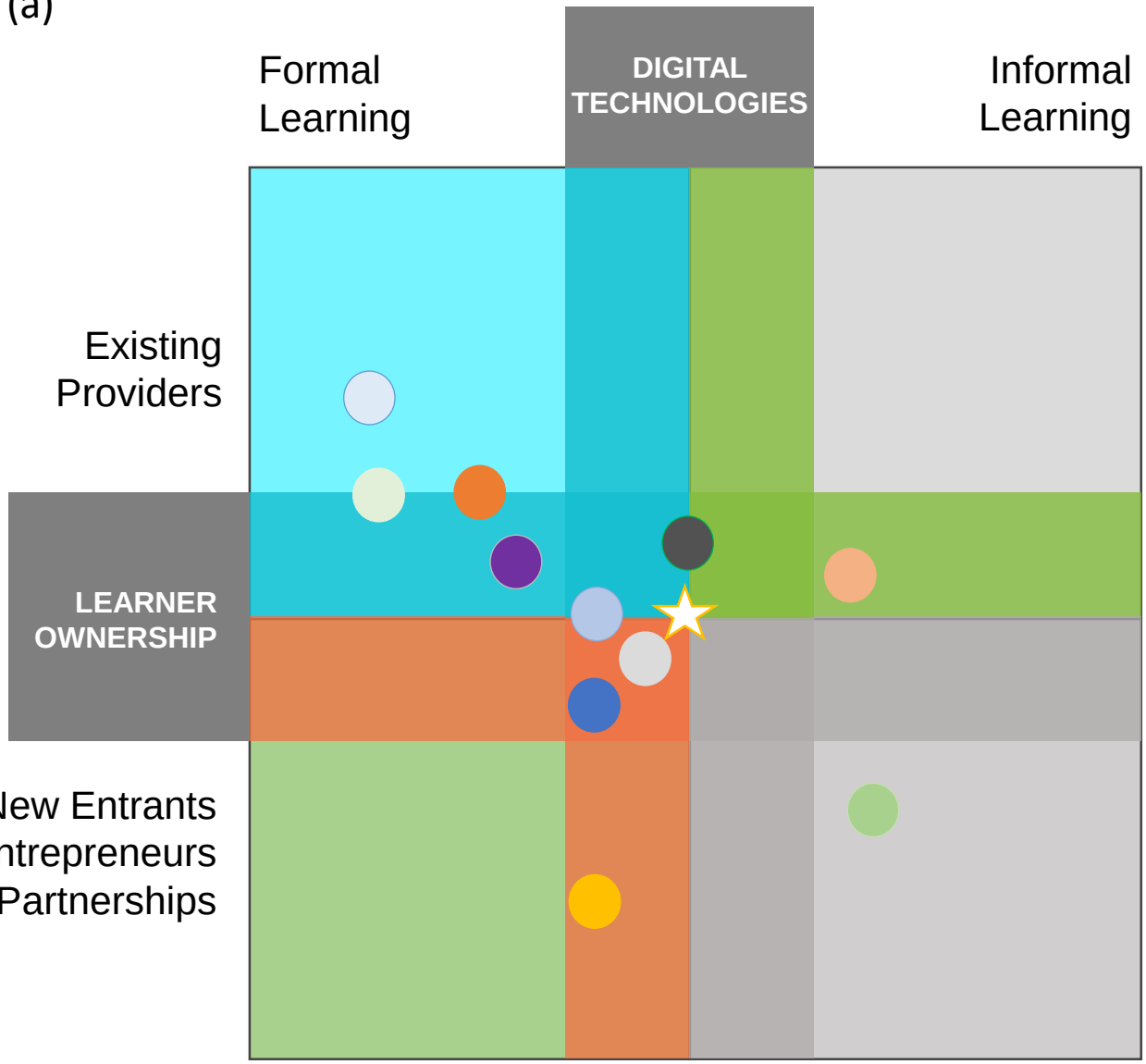
Review framework

This framework is adapted from the Innovation Ecosystem 2x2 which helps us to map the contribution of different players within a system. In this instance, we have populated it with examples from this collection of assessment tools.

The aims of using the grid include: to identify where there are opportunities left uncovered; whether a particular player is engaging with one or more parts of the system; and whether it is engaging with key drivers of new forms of learning: digital technologies and learner ownership.

In our workshop, group can test out and edit the axes to focus on different purposes of assessment tools, for example, instead of formal or informal learning, we might distinguish between whether a tool assesses cognitive or noncognitive skills.

(a)



Scottish Highers
Hong Kong Diploma
Amplify – early literacy
NH competency frameworks
NY Performance consortium
GRIT scale
Torrance - creative thinking
Mission Skills Assessment
DiscoTests
Queensland Performance Ass.
Singapore Project Work