Chapter 2
Assessment of Twenty-First Century Skills: The Issue of Authenticity

Esther Care and Helyn Kim

Abstract Writing skills are assessed through writing tests, typing skills are assessed through typing; how do we assess critical thinking or collaboration? As interest in twenty-first century skills increases globally, and as skills goals are explicitly adopted into curricula, the inadequacy of our knowledge of how these skills develop becomes increasingly problematic. These goals reflect human processes, both cognitive and social, and this challenges many current assessment approaches. To highlight some of the issues associated with assessment of twenty-first century skills, a review of a sample of assessment tools was undertaken. The review provides some insights both into how far we have come as well as how far we have to go. The diversity of the tools and evaluation of these against authenticity dimensions highlights the challenges not only in design of assessment but in how teachers might design classroom learning experiences that facilitate development of twenty-first century skills.

Introduction

There is global recognition of the need for students to develop a broader set of skills during the years of formal education than has traditionally been the case. Although recognition of importance of work-ready skills has long been endorsed, it is relatively recent that calls for their development have moved from a strongly vocational stance (e.g., Brewer 2013) to an education for both work and life perspective (e.g., Pellegrino and Hilton 2012).

In many countries, education ministries commit to goals such as developing "the whole person", characterised by sets of values, ethics, and attitudes aligned with national identity, as well as developing students' social-emotional characteristics and cognitive skills. Introduction of twenty-first century curricula requires knowledge and understanding of how the aspirations in mission statements

E. Care (✉) • H. Kim
Brookings Institution, Washington, DC, USA
e-mail: ecare@brookings.edu

E. Care et al. (eds.), *Assessment and Teaching of 21st Century Skills*,
Educational Assessment in an Information Age,
DOI 10.1007/978-3-319-65368-6_2

translate into the particulars of what students need to learn and know how to do, and of what teachers need to teach and know how to assess. Given that a primary justification for assessment is to improve student educational outcomes, information from assessment must be aligned with the purposes to which it will be applied (Almond 2010).

Assessment is often dichotomised across summative and formative functions. Another function is as a driver of teaching and learning. For example, the fact of assessing particular domains is sometimes seen as signalling that the domain is valued by the system (Schwartz et al. 2011) particularly to teachers. Where the main interest is in stimulating learning and competency development (Birenbaum 1996), authenticity of the assessment is pre-eminent since we are interested in predictive capacity of results. Taking four "21st century skills" that have recently been identified by country education systems as valued (Care and Luo 2016; Care et al. 2016a), in this chapter we highlight authenticity issues associated with their assessment through exploration of tools designed to measure them. The four skills are problem solving, collaborative problem solving, computer and information literacy, and global citizenship.

Complex Nature of Skills

Some skills, such as problem solving, might be seen as uni-dimensional in the sense that just one main type of contributing factor – cognitive skills – describes them, although multiple processes contribute to them. Other skills are clearly multidimensional by virtue of drawing on qualitatively different skills. Collaborative problem solving is a case in point. It combines the two broad domains of social and cognitive skills, and within these, calls on the skills of collaboration and problem solving. In turn, each of these is comprised of more finely delineated subskills such as responding, organising information, and so on. Such skills might be referred to as complex skillsets (Care et al. 2016a; Scoular et al. 2017) or complex constructs (Ercikan and Oliveri 2016). Another complex skillset is global citizenship, which is hypothesised to draw on social and cognitive capacities as well as values, knowledge and attitudes. Such complex skillsets pose additional challenges for measurement due to the difficulty of identifying the degree to which each subskill might contribute unique variance, or the degree to which demonstration of one subskill might depend on reaching some hurdle level of competence in another.

The research phase of the Assessment and Teaching of 21st Century Skills (ATC21S) project highlighted two complex skillsets – collaborative problem solving and digital literacy in social networks. Through its exploration of these, one of the project's major contributions was clarification of our understandings of these skills. This understanding culminated in the development of tools for assessment and consideration of curricular and pedagogical implications. The research contributed in particular to global perceptions of the nature of collaborative problem solving (OECD 2013), as well as to discussion about innovative forms of assessment.

The assessment approach taken by ATC21S was decided upon in response both to the nature of the complex skillsets of interest, and the affordances of online data capture. The use of the ATC21S tools has been largely confined to research studies and has provided valuable insights about the degree to which online data capture of student action can inform estimates of student performance across both social and cognitive activities (Care and Griffin 2017).

The Assessment Challenge for Twenty-First Century Skills

Demonstration of skills or competencies is through behaviours which we hypothesise are accounted for by latent traits. ATC21S therefore targeted behaviours for capture in order to draw inferences about these traits. This approach is quite different from targeting individual's perceptions about their latent traits (as demonstrated through self-report techniques), or knowledge or reasoning capacities (as demonstrated through correct/incorrect responses to test items). And here lies one of the challenges for assessment.

In comparison to the educational assessment of content-based knowledge, assessment of twenty-first century skills is in its infancy. To date, there has been little attention paid to construct validation of assessments in the classroom, or to predictive validity based on evidence of the generalisability of skills-based learning. Challenges in assessing twenty-first century skills lie in our lack of comprehensive understanding of the nature and development of the skills, about their multidimensionality, and about how to partition variance in behaviour that is attributable to knowledge, or attributable to skill.

These issues are key for psychometricians in developing standardised instruments as well as for classroom teachers in developing classroom based tasks. Critical to skills domains is the assumption of developmental trajectories (Gee 2010). Knowledge of the skills requires not only identification of contributing subskills, but also evidence of how these individually and together progress, from simple to advanced. This explains the need to design tasks that require demonstration of skills at increasing difficulty levels.

An issue in design of assessments of twenty-first century skills is the degree to which assessment tasks actually stimulate the processes that indicate the targeted construct and provide a facility for their capture. To stimulate them, it is essential that the assessment design, as much as possible, mirrors the authentic demands of the situation that provoke behaviors associated with the targeted skill (Care et al. 2016a, b). Ercikan and Oliveri (2016) address this challenge by proposing to acknowledge the complexity of the construct and systematically align tasks with different elements of the construct. This raises questions about whether the construct itself is being assessed, or merely some of its components. Of interest is whether an assessment takes a form that can capture the true nature of the skills and report on this in a way that represents the skill in varying degrees of competency.

A Focus on Authenticity

There has been a rapid spread in the twenty-first century skills phenomenon in formal education (Care and Luo 2016). The intention of explicit focus on twenty-first century skills in education is that students will develop the capacity to apply these skills to real life situations. Hence, assessment tasks should be authentic (Gulikers et al. 2004) – that is, reflect the characteristics of long-term professional work and life behaviours. This means that assessment tools must be designed to capture the cognitive and social processes rather than factual knowledge. Authenticity does not guarantee construct validity – whether an assessment actually measures what it purports to measure – but can contribute evidence to support it. This evidence may be derived from tasks that reflect the competency of interest, represent a realistic application of the competency, and reflect the cognitive and social processes that contribute to the behaviour in real life.

In 1996, USA's National Research Council (NRC) called for assessment to support educational reform for the twenty-first century. The NRC proposed more focus on learning processes as opposed to learning outcomes; more targeted assessment as implied by a focus on what learners understand and can do; and rich or authentic knowledge and skills. These goals are aligned not only with a competencies approach in education, but to principles of formative assessment. Student-centred pedagogies that rest on formative assessment are well aligned with concepts of skills development. A majority of twenty-first century skills are demonstrated through actions, and therefore require an interactive style of pedagogy as opposed to transmission paradigms. Accordingly, assessment needs to attend to actions and behaviours, or enable inferences to be drawn from these. Central to the rationale for twenty-first century skills education is the degree to which students can develop skills that can be applied across different contexts (Blomeke et al. 2015); the whole point is to develop in students the capacity to generalise, to adapt and to apply. How can assessment capture these applications?

This brief review examines the degree to which selected tools are consistent with five characteristics of authentic assessment defined by Gulikers et al. (2004). This in no way competes with current views on validity as represented in standards for educational and psychological assessment (e.g. AERA/APA/NCME 2014), but is complementary. As pointed out by Pellegrino et al. (2016) "an assessment is a tool designed to observe students' behavior and produce data that can be used to draw reasonable inferences about what students know" (p. 5). This definition clearly addresses tools designed for educational purposes. Pellegrino et al.'s (2016) interest in instructional validity is consistent with concerns about authentic assessment (e.g., Wiggins 1989; Gulikers et al. 2004).

Review of Selected Tools

Gulikers et al. (2004) state that authenticity lies in "an assessment requiring students to use the same competencies, or combinations of knowledge, skills, and attitudes, that they need to apply in the criterion situation in professional life. The level of authenticity of an assessment is thus defined by its degree of resemblance to the criterion situation" (p. 69). Authenticity also needs to reflect a learning approach from students' early basic skills through to those in final years of secondary school where behaviours are displayed that are more recognisable as the mature skills. Assessments need to reflect this progression.

Gulikers et al.'s (2004) five dimensions of authentic assessment are:

(a) An authentic task presents as a set of activities that emulate professional practice
(b) The physical context reflects the way the competencies will be applied in professional practice
(c) The social processes (if these are relevant) will reflect those applied in the real situation
(d) The product or performance mirrors a real life one, permits inferences about the underlying construct, includes multiple indicators, and is available to others for review
(e) Criteria identify what is valued, and standards indicate levels of performance expected.

One aspect of Gulikers et al. (2004) model is including the student perspective on relevance of task. Although this information might well be collected during development of assessments, through cognitive laboratories or interviews, it is rarely included in test manuals of large scale assessments. Evidence addressing this in the review is therefore slight.

For the four selected twenty-first century skills (problem solving, collaborative problem solving, communication and information literacy, and global citizenship), one measure of each was chosen to illustrate and consider the authenticity of current assessments from the perspective of these five dimensions. The search for example assessments was conducted systematically. First, specific key words and phrases were entered into search engines (google, google scholar, bing). These key words and phrases included: "assessments of 21st century skills", "21st century skills", "large scale assessments", "key competencies", "collaboration", "problem solving", "information and communication literacy", "technology", "global citizenship", or some combinations of these words and phrases. Based on these searches, reports and articles were accessed and explored to gather a pool of assessments of twenty-first century skills.

In order to select just one assessment tool of each skill, the database of tools was successively refined. Initially, two criteria were used: intended for use at large scale for school populations; and availability of technical and/or research information. Tools were then filtered out if discontinued as of October 2016; if in fact were second or third party rating tools; or were measures that assess course knowledge for academic qualifications or those that are part of program-based toolkits or badged programs. For example, the Pearson Edexcel International GCSE Global Citizenship[1] exam comprises an externally-assessed paper, which is given after

[1] http://qualifications.pearson.com/en/qualifications/edexcel-international-gcses-and-edexcel-certificates/international-gcse-global-citizenship-2017.html.

Table 1.1 Tools in the context of five dimensions of authenticity

		Five-dimensional framework for authentic assessment				
21st century skill	Tool	Assessment task	Physical context	Social context	Assessment result or form	Assessment criteria
Problem solving	MicroDYN (Funke 2001; Greiff and Funke 2009)	Items require participants to control dynamic and interactive simulated systems, which are designed to mirror what may occur in real-life, and require the integration of knowledge, skills, and attitudes to complete the task. Example task is the virtual chemical laboratory, where students have to understand causal relations among chemical substances and elements and build models and make predictions based on knowledge gathered while exploring the system.	Participants receive 8–12 items lasting about 6 min each, totalling roughly 1 h. There is high fidelity in terms of how close the system imitates reality, in terms of the underlying processes (e.g., retrieve information and apply it to make predictions and build models).	Not specifically mentioned; but task is completed individually.	Participants generate solutions through a performance-based task; Includes items with different underlying structures and difficulty levels.	Criteria are embedded within items based on characteristics valued and used in real life; some expectations are made explicit, whereas others are not.
Collaborative problem solving	ATC 21 collaborative problem solving (Care and Griffin 2014)	Task requires collaborative effort and integration of cognitive and social skills to reach a solution, resembling the complex nature of the construct. Students' ownership of task and processes are rewarded, rather than the solution. Example is Olive Oil task, which reflects Tower of Hanoi-like problem.	Tasks simulate scenarios within a technology context and mimic the resources and tools available in real-life, as well as the asymmetric nature of solving problems.	Students have differing resources, tools, and information, and must work together to establish the series of steps to solve a problem (not face to face interactions).	Result relies on actions and processes captured in logfiles; multiple tasks are available to be "bundled", in order to capture the construct comprehensively.	Criteria and standards are pre-determined and task-specific. Logfiles are coded and mapped onto cognitive and social dimensions that are linked to specific competence levels.

Computer and Information Literacy	IEA International Computer and Literacy Study (Fraillon et al. 2015)	Tasks require integration of knowledge and skills in a simulated environment that mirrors real life situations. Example is working with collaborators to plan a design of a new garden for their school. Final product is an information sheet that explains the design and convinces peers to vote to use that design.	Computer-based task simulates real-life situations; Estimated time component for each module is between 20 and 30 min.	Incorporates social processes due to requirement for collaboration, which occurs entirely within the test platform (not face to face interactions).	Results are in the form of product and performance, such as creating a design plan.	Criteria and standards are in the form of an achievement scale that maps onto proficiency levels, which describe skills and knowledge that are expected at each level.
Global citizenship	Southeast Asian primary learning metrics global citizenship domain (Parker and Fraillon 2016)	Items are student self-reports on attitudes and values, and to a certain extent, behaviours and skills related to global citizenship issues.	Student questionnaire does not capture how knowledge, skills, and attitudes are used in real-life situations; low fidelity since environment (self-rating) does not imitate reality; 20–30 min to complete the assesment	Not specifically mentioned	Items assess attitudes and values in the form of a student questionnaire.	Criteria and standards are not specified.

completion of a 2-year course for teaching in international schools or undertaking community action on a global issue. Despite the focus on global citizenship (Pearson 2017), such a tool would not be included in the review.

The four large scale assessments selected were developed for summative purposes in the first instance. As shown in Table 1.1, they are assessments of: problem solving (MicroDYN; Greiff and Funke 2009); collaborative problem solving (ATC21S; Care and Griffin 2014); communication and information literacy (IEA International Computer and Literacy Study [ICILS]; Fraillon et al. 2015); and global citizenship (South East Asian Primary Learning Metrics; Parker and Fraillon 2016).

Problem Solving

Problem solving involves being able to negotiate complex and dynamically changing environments and situations successfully by drawing on behavioural patterns to reach a desired goal (Funke 2003; Greiff et al. 2013). More specifically, dynamic or complex problem solving has been defined as "the successful interaction with task environments that are dynamic (i.e., change as a function of user's intervention and/or as a function of time) and in which some, if not all, of the environment's regularities can only be revealed by successful exploration and integration of the information gained in that process" (Buchner 1995, p. 14). Therefore, someone who is successful at problem solving is able to interact with the task environment and adapt to the dynamic nature of these environments in order to collect information; integrate and structure information in a meaningful way; and effectively apply the acquired knowledge to make predictions and solve the problem at hand (Dörner 1986; Mayer and Wittrock 2006).

Due to the complexity of the construct, measuring problem solving is also complex. Assessments of problem solving depend on the flexibility of tools and platforms to capture problem solving abilities in dynamically changing contexts (Greiff et al. 2013). Hence, computer-based performance assessments are well equipped to capture the acquisition and application of knowledge to solve complex problems. MicroDYN (Funke 2001; Greiff and Funke 2009), a computer-based assessment of complex problem solving was chosen as an example to examine authenticity. MicroDYN is based on a framework (Funke 2001) in which inputs affect outputs. For instance, increasing an input variable might result in a decrease or an increase in one or more output variables in the system (Greiff et al. 2012). The user interacts with and navigates through an unfamiliar system or situation which mirrors problem solving in real-life settings. Participants are prompted to detect causal relations and control systems that are presented. There are three stages underlying each MicroDYN item that align with three aspects of problem solving: (1) exploration, where participants can explore the system freely with no restrictions, and use strategies to retrieve information about the system; (2) drawing mental models, where participants draw the assumed connections between variables as they understand it to be; and (3) forecasting phase, where participants attempt to achieve target values in the output variables by entering correct values in the input variables within a fixed number of steps. This is the stage where practical application of acquired knowledge from the previous stages is assessed. Eight to twelve independent items are presented to the participants in dynamic and interactive situations.

The assessment task is authentic in that it confronts students with situations that mirror what occurs in professional practice – meaningful and relevant and requires the knowledge, skills, and attitudes to complete the task. For example, one MicroDYN test is composed of 11 independent tasks and 2 trial tasks that are embedded in the context of a virtual chemical laboratory. The students are presented with chemical substances and elements and need to understand their interrelations to build models and forecast. In addition, there appears to be autonomy in the exploration phase where students can explore the system freely. The fact that items are designed to activate minimal prior knowledge provides support for authenticity – only the specific knowledge gathered during the task is relevant. When problems rely on prior knowledge and specific content, solutions tend to be routinely available, which detracts from the essence of complex problem solving, to which dynamic interaction with an unknown environment is key (Greiff et al. 2012).

The physical context of the MicroDYN items is authentic in that it reflects the way knowledge, skills, and attitudes will be used in real-life situations, although it can be argued that the physical context is less authentic – being a test platform. Regardless, students need to use strategies to retrieve information about the system, and then integrate and apply that information to draw connections between variables in the model, and finally achieve target values based on hypotheses about how inputs and outputs are related. There is high fidelity in terms of how the MicroDYN systems imitate reality. The time limit of five to six minutes per item may seem counter to the amount of time that is available to solve problems in real life; however, according to the developers of the MicroDYN approach, "a short time on task is ecologically valid because successful interaction with unknown task environments in real life seldom lasts longer than a few minutes (e.g., buying a ticket at a new vending machine)" (Greiff et al. 2012, p. 193). In terms of the social processes of the MicroDYN approach, there is no specific mention of the social context; the test is designed for individual completion rather than in collaboration with others, reflecting many real-life situations. The assessment result is in the form of information retrieval through exploring the simulated systems, building models by drawing connections between variables, and then applying the acquired knowledge by controlling and meeting target values. Authenticity of assessment result lies with students demonstrating their competencies by generating solutions through a performance-based task and by engaging with items with different underlying structures and difficulty levels.

Finally, the criteria for an authentic assessment are referred to the dimensions of the framework. Embedded within the items are characteristics valued and used in real life, including "the ability to use knowledge, to plan actions, and to react to dynamic changes" (Greiff et al. 2012, p. 195). Some expectations are made transparent and explicit beforehand. For instance, during the information retrieval and model building stages, students are explicitly told that they do not have to achieve specific target values as they are navigating and gathering knowledge about the system but that they will be asked to do so later on. How items are scored varies and is not made explicit to students beforehand.

Collaborative Problem Solving

Collaborative problem solving (CPS) is defined as a complex skill that requires both cognitive and social processes (Care and Griffin 2014). CPS has been hypothesised as consisting of five strands of participation, perspective taking, social and task regulation, and knowledge building, and is brought to bear when the ability or resources of a single person is not enough to solve a problem. Individuals need to be able to combine various resources and skills when confronted with a complex problem (Hesse et al. 2015).

Putting aside arguments (Scoular et al. 2017; Rosen and Foltz 2014; Rosen 2015) concerning authenticity of assessment of collaborative problem solving that involves agents as opposed to people, of interest in this discussion is the degree to which the problem solving environment mirrors real life freedom of movement in the exercise of cognitive and social competencies by a pair of problem solvers. Any online platform imposes constraints on freedom of movement, but can vary through presentation of well-defined versus ill-defined problems.

The ATC21S Collaborative Problem Solving environment (Care et al. 2015, 2016a, b) was used to examine the authenticity of the assessment. The assessment is online with eleven tasks that are designed to capture human to human collaborative problem solving. Some tasks are asymmetric, meaning that each student engaging with the tasks has access to different, but critical, information to solve the problem, and exemplified by the Olive Oil task which reflects the Tower of Hanoi style of problem. As the students work on solving the problem, actions, chat events, and combinations of both are captured in a logstream file for coding and scoring of CPS competence based on the hypothesised cognitive and social underpinnings. The scores are then used to indicate the student's level of CPS competence.

ATC21S CPS presents an authentic assessment task that requires the integration of cognitive and social skills. For example, one indicator is that the student "uses understanding of cause and effect to develop a plan", which corresponds to the element of knowing the "if-then" rule, which in turn, captures the broader strand of knowledge building (Care and Griffin 2014). The task illustrates the complexity of the construct, involving unstructured exploration of the problem space, as well as arriving at a solution in multiple ways. The assessment rewards processes enacted rather than the actual solution (Adams et al. 2015). The CPS tasks allow students to take ownership of the process of reaching a solution and are designed to capture the transferable or generalizable skills involved in the types of problems that require real life collaborative effort. No pathways are pre-defined except insofar as students must move forward from one screen to another to complete the tasks.

The CPS tasks are intended to simulate scenarios that may occur in learning and teaching environments as students work together to solve complex problems. The physical context is a technology context but does not detract from the fact that the tasks elicit the cognitive and social processes involved in CPS, including problem analysis, planning and executing, and awareness of and ability to adapt to whoever is the partner. The task context mimics the resources and tools available in real-life situations; and the way skills will be used in professional settings. The fact that some tasks are asymmetric in nature, with each student coming to the problem with different resources, is similar to real life problems. The social context supports the authenticity of the assessment as students who are provided with different resources, tools, and information, must work together to establish how the tools function, whether they are necessary to solve the problem, and the series of steps to follow. Assessment result and form rely on the actions and processes of the students, which are captured using logfiles. This type of result and form are analogous to those which occur in real life in that it depicts the processes and actions that are undertaken to solve complex problems in professional capacities. The developers recognise that any one task may not necessarily capture the construct in a comprehensive way (Care et al. 2016b). Therefore, multiple tasks are available that can be bundled to "provide a comprehensive sampling across the construct, as well as the capacity to assess students at different levels of competence" (p. 14). The tasks do not require students to defend their solution, as may occur in real-life settings, although the free form chat could promote that activity. Finally, the criteria and standards are predetermined and task-specific. The coded indicators based on the actions and chat events gathered from logfiles are mapped onto cognitive and social dimensions, which are identified across competency levels.

Although ATC21S CPS has received major psychometric attention, the challenge of capturing this complex multi-dimensional construct remains. Technical information is presented in Scoular et al. (2017). Primary issues relate to lack of evidence of construct validity evidence beyond model-fitting techniques, difficulty in capturing social processes in an online environment, and capacity of an online platform to provide a sufficiently unstructured environment in which sophisticated levels of the skill might be demonstrated.

Computer and Information Literacy

As global economies seek to maintain productivity and embrace technological advances, equipping students with information and communication technology (ICT) and digital literacy skills is important for their full participation and success in today's information-rich, technology-driven society (Kozma 2011). From the basic ability to use computers and other technology devices individuals need to process, evaluate, and retrieve information (Catts and Lau 2008), participate in social networks to create and share knowledge, and to use and produce digital media.

The International Association of the Evaluation of Educational Achievement's (IEA) International Computer and Information Literacy Study (ICILS) (Fraillon et al. 2015) is used as an example to consider authenticity of assessment. Computer and information literacy (CIL) comprises two overarching categories or strands. The first strand focuses on collecting and managing information. The second strand focuses on producing and exchanging information, including transforming, creating, sharing, and using. CIL is assessed through four computer-based assessment modules that follow a linear narrative structure using a combination of purpose-built applications and existing software. Students navigate the system, as well as complete questions and tasks which are delivered in 30-minute modules. Students complete two of four available modules, with each module including a series of smaller five to eight tasks with a total task time of 15–20 min. Three task types include: (1) information-based response tasks, which use computer technology to deliver pencil-and-paper-like questions using multiple-choice, constructed-response, or drag-and-drop- response formats; (2) interactive simulations or universal applications to complete an action, such as navigating through a menu structure, and capturing "correct" responses; and (3) authoring tasks, in which students modify or create information products using software applications. The test items are automatically scored, and the score is placed on a CIL achievement scale corresponding to proficiency levels (Fraillon et al. 2014).

Specifically, ICILS defines CIL as "an individual's ability to use computers to investigate, create, and communicate in order to participate effectively at home, at school, in the workplace, and in society" (Fraillon et al. 2015, p. 17). ICILS provides authentic assessment tasks that require students to integrate their knowledge and skills in a simulated environment that mirrors the kinds of tasks they may face in real-life situations. For example, one module requires students to work with a group of collaborators to plan the design of a new garden area for their school. The final product is a student-generated information sheet that explains the garden design, as well as creates support for that particular design, so that their classmates will vote to use the design. Attitudes are assessed separately through a student questionnaire. The physical context of the test is authentic in that the computer-based task simulates real-life scenarios. The garden design example simulates computer-based professional landscape design technologies to communicate information. However, there is an estimated time component for each of the modules of about 20–30 min, whereas in professional activities, this assignment would presumably involve a longer period. To this extent, the task may not reflect the real-life complexity. As for the social context, the tasks incorporate the social processes that are drawn upon in practice. In real life, architects and designers may work individually; but more often than not, a larger project would require multiple people with differing expertise working together to create the final design product. In ICILS, the collaboration occurs entirely within the test platform rather than through face-to-face interactions. Assessment result and form rely on products and performance by students, which are similar in nature to the kinds of products that professionals may be asked to generate in professional settings (i.e., a design plan). Finally, the criteria and standards are specified in the form of an achievement scale that maps onto proficiency levels. The proficiency levels describe the kinds of skills and knowledge that are valued and expected at the various levels. The criteria are pre-determined but it is unclear whether students have access to the descriptions beforehand to guide their learning (Sluijsmans 2002). How the scoring of the items locates where the student may fall along the achievement scale is similarly unclear.

Global Citizenship

The significance of global citizenship education (GCED) in promoting sustainable development, equity, and inclusive societies is well-recognized (United Nations Educational, Scientific and Cultural Organization, UNESCO 2014). Global citizenship can be broadly described as a sense of collective identity, belonging to a global community, with the implication that people are connected in multiple ways to each other and to their environments (UNESCO 2014).

The Southeast Asian Primary Learning Metrics (SEA-PLM) Global Citizenship Domain Assessment (Parker and Fraillon 2016), an assessment of the attitudes and values related to global citizenship, is provided as an example to examine authenticity. SEA-PLM focuses on the attitudes and values (e.g., feeling, sensing, valuing, believing) related to global citizenship and "reflects the dispositions that can lead to deeper engagement with global citizenship in the later years of schooling" (p. 6). A student questionnaire, adopting Likert scale response options, addresses attitudes toward global citizenship systems, issues and dynamics; citizenship awareness and identity; and global citizenship engagement. A few items also ask about students' experience of activities related to global citizenship, such as presenting ideas or leadership, to capture behavioural aspects of the construct. Items target awareness of diversity in society, knowledge of concepts of citizenship including "good citizens" and "global citizens", knowledge of benefits and consequences of personal and collective civic engagement, attitudes toward the value of learning about global citizenship, self-reported behaviour associated with global citizenship, and attitude and behavioural intentions with regard to protecting the environment (Parker and Fraillon 2016). A teacher questionnaire is forthcoming.

The development of the SEA-PLM Global Citizenship Domain assessment is grounded in the following working definition of global citizenship:

> *Global citizens appreciate and understand the interconnectedness of all life on the planet. They act and relate to others with this understanding to make the world a more peaceful, just, safe and sustainable place* (Parker and Fraillon 2016, p. 5).

The authenticity of the assessment task as it currently stands is questionable, first and foremost because a self-rating student questionnaire is used. Research suggests there are three major competencies related to global citizenship: cognitive aspects (i.e., knowledge acquired about global structures, systems, and issues); attitudes and values about global citizenship concepts (e.g., appreciation of diversity, equity, non-violence, social justice); and behaviours and skills involved in participating in activities that create “positive change and foster social participation” (Parker and Fraillon 2016, p. 5). The student SEA-PLM global citizenship assessment focuses primarily on the dimensions of attitudes and values, and less on behaviours and skills. An important dimension of authenticity is that assessments are not atomistic (Gulikers et al. 2006), and that tasks reflect underlying dimensions according to performance, as opposed to what respondents *think* about what they would do, or about their own traits. According to the developers of the assessment, factors including time, age and grade level, contributed to decisions upon which components to focus. The initial targeting of Grade 5 students for the field trial, as well as the limited time available for the assessment, influenced the decision to focus mainly on attitudes and values. As a result, the complexity of the construct is not captured, nor is ownership of the task reflected for students as they are not engaging in global citizenship-related activities.

Although the context of the assessment is not specifically identified, the student questionnaire format means that the physical context does not reflect the way knowledge, skills, and attitudes will be used in real-life settings. The assessment has low fidelity, since the environment does not closely imitate reality (Alessi 1988). The method itself of assessing global citizenship detracts from the capacity of the tool to capture student ability to use global citizenship competencies. Relatedly, the assessment is given in 20–30 min. Whether global citizenship responses could be generated within such a restricted time period is questionable. Similar to the physical context, the social context is not mentioned. Global citizenship includes a sense of interconnectedness of citizens around the globe. This implies the importance of the social context that fosters a sense of belonging to the global community. This is not captured by the assessment. In terms of the assessment form and result, indicators tap attitudes and values rather than the full construct. Although questionnaire items ask about opportunities students have had for active participatory engagement, this cannot reflect a competency; other methods such as a product or performance may be required to demonstrate mastery (Darling-Hammond and Snyder 2000). Finally, criteria and standards are not specified. To note, the SEA-PLM global citizenship domain survey remains under development, having gone through field trials in 2015–2016 (Parker and Fraillon 2016).

The approach adopted by SEA-PLM reflects traditional methods of measuring attitudes and values through self-rating surveys. Taking into consideration the frameworks for global competence specifying that knowledge, skills, attitudes and values lead to competencies and action (Ramos and Schleicher 2016), it is clear that SEA-PLM has followed the line that assessment of global competence and citizenship can be achieved by measurement of these predictors, rather than targeting the competencies and actions. To date, although knowledge and attitudes are clearly targeted, attention to skills is not so clear.

Perhaps reflecting this state of the art and perspective, PISA 2018 global competency measurement will rely on cognitive items that reflect knowledge, perspective-taking and analytical and critical thinking components. Along with multiple choice items, OECD proposes use of critical incident case studies which prompt open-ended responses to be scored with use of rubrics. Self-report on skills and attitudes will be used for reporting at country or sub-population level. These decisions are the clearest communication by the OECD assessment community that authentic assessment of social competencies underpinned by values, attitudes, and beliefs, still eludes us (Ramos and Schleicher 2016).

Discussion

Authenticity informs validity. The emphasis on authenticity in this chapter is to draw attention to how modern assessments stimulate, capture, score and evaluate in ways that might contribute supporting evidence to validity. Although the focus of the chapter is assessment, the authenticity demand within the classroom teaching and learning context is analogous – and equally demanding. The sample of assessment tools demonstrates both traditional and innovative approaches to assessment of twenty-first century skills. They range from Likert scale self-ratings of attitudes and values to rich computer-based task environments where students can display a repertoire of skills. Are we actually capturing indications of the skills of interest? The majority of published technical information on the tools reviewed consists of explorations of the internal structures of the tools. This is not sufficient to inform judgements about the degree to which we can rely on assessment data to understand student capabilities or readiness to learn.

The ICT literacy tasks stand out in the authenticity stakes, given that the assessment mode is firmly situated within the operating environment required for real life enactment of the skills. The examples demonstrate use of rich task environments, concentrating on how the individual accesses and uses technology-based artefacts.

Assessment of problem solving illustrates innovative approaches to measurement of the skills through capturing processes using twenty-first century technologies. With strong reliance on online facilities, assessment of problem solving still uses traditional multiple choice options, but has also moved into logging student progress through rich online tasks, and attempting to interpret the activity data trail left by the student through engaging with the task environments. Capturing student activity, coding it, and trying to make sense of it, is the state of the art.

The main skill area in which boundaries have been pushed is collaborative problem solving. The lure of the cognitive, or problem solving, aspects of the construct, with which much progress has been made, makes both closer and more tantalising the challenge of capturing the social processes that are brought to bear in a collaborative environment (e.g. von Davier et al. 2017).

The area that is not associated with state of the art assessment approaches is global citizenship. Although often referred to as a skill, in fact this includes (1) processes brought to bear in decision-making, critical thinking, and problem solving; (2) social processes including communication, which are intrinsic to resolving wicked problems; and (3) values and attitudes. Approaches to assessment of these latter continue to rely on approaches derived from psychological measures including self-rating of characteristics, attitudes or values. Lundberg's (2015) characterisation of 'metrics' of non-cognitive skills as falling across three categories – self-assessments, parent/teacher reports, and extrinsic administrative indicators holds true for this competency.

The defining characteristic of a twenty-first century skill is that an individual or group of individuals can bring that competency to bear in and across new situations including those associated with or within technology environments. This very characteristic is what challenges assessment. How to measure the non-routine is what confronts us. With well-known skills such as literacy, assessment tools are able to capture both early and more developed competencies. With complex skills that are less well-known, our current assessment technologies also appear to be able to capture some early subskills, but are less capable of capturing the more developed competencies since these latter are exercised in free ranging ways and in environments that do not offer the opportunity for reliable data capture, coding or scoring given our current technological progress. For the early forms of competencies, we make assumptions that the behaviours sampled are predicted by the complex skill which will account for the individual's performance in real-life situations. Our inability to capture the more developed competencies puts at question our capacity to measure the full range of skill. This is a crucial validity threat to instruments. Another important consideration concerns the degree to which what is measured reflects all aspects of the competency (Soland et al. 2013). For example, measurement of collaborative problem solving, both by PISA OECD and by ATC21S, has found eliciting objective measurement of subskills of the social domains elusive while other aspects of the construct are captured reliably.

Against a backdrop of consensus globally about the importance of twenty-first century skills for equipping future generations to live constructively, implementation by education systems through reformed curricula, dynamic pedagogies, and innovative and aligned forms of assessment lags behind. The findings from the Network on Education Quality Monitoring in the Asia-Pacific (NEQMAP) study of the assessment of transversal competencies (Care and Luo 2016) across nine countries in Asia Pacific demonstrates that classroom assessment tools are no more sophisticated or prolific than are large scale assessment tools. Lack of clear guidelines about how to assess from national system level is reflected at school level, and signified by calls from teachers for more guidance about the nature of twenty-first century skills, how to teach, and how to assess them. Recent mapping of national education mission statements with a focus on twenty-first century skills (Care et al. 2016a) demonstrates the lag across aspiration and implementation in the actual curriculum. From the brief review in this chapter, similarly is seen a scattering of assessments that are intended to measure these skills but vary from reliance on traditional methods to testing the possibilities of innovative methods of data capture, interpretation, and use.

As pointed out by Csapó et al. (2012), an issue that has confronted assessment experts is the constraint placed on capturing a construct where paper and pencil is the major capture medium. Notwithstanding that electronic media expand the opportunities for capture, they do not solve the other major issue – that of ensuring that what is captured is interpretable. The issue is well demonstrated by this review of the assessment tools – that the capture medium has expanded widely, but still falls short of providing an authentic environment in which the skills can be freely exercised yet reliably captured. In addition, the majority of effort has been dedicated to crafting the opportunities, and checking internal indications of psychometric robustness, rather than looking to concurrent or face validation opportunities. Ercigan and Oliveri (2016) consider three sets of factors that relate to validity in the consideration of assessment of twenty-first century skills: construct complexity and how this influences task design in the context of generalisability; the use of empirical data that can provide evidence that relates to student performance in real life; and cross-cultural and context issues. Blomeke et al. (2015) focus on two validation approaches: first, a model fitting approach comprising hypothesis, followed by analysis to see if the data fit; and second, a "real-life" approach in which the effort is to obtain measures as closely related to criterion performance as possible. The four tools reviewed here demonstrate varied levels of authenticity, which go some way to addressing criterion performance, but as of yet, their predictive capacity is not evident.

There is no doubt that our capacity for data capture of assessment transactions has taken great strides. The challenge remains first, to provide stimulus environments for the transactions that are themselves aligned with the nature of what is to be measured; and second, to capture the indicators of skills in a reliable manner that is interpretable in terms of competence levels. These challenges have clear implications for the demands on the teaching and learning environment, and the degree to which the teacher can create the context in which the same skills can be nurtured.

Acknowledgement The authors wish to thank Scott Paris of the Australian Council for Educational Research for extensive and constructive comments to earlier versions of this manuscript.

References

Adams, R., Vista, A., Scoular, C., Awwal, N., Griffin, P., & Care, E. (2015). Automatic coding procedures. In P. Griffin & E. Care (Eds.), *Assessment and teaching of 21st century skills: Methods and approach*. Dordrecht: Springer.

Alessi, S. M. (1988). Fidelity in the design of instructional simulations. *Journal of Computer-Based Instruction, 15*, 40–47.

Almond, R. G. (2010). Using evidence centered design to think about assessments. In V. J. Shute & B. J. Becker (Eds.), *Innovative assessment for the 21st century* (pp. 75–100). Boston: Springer Science+Business Media. doi:10.1007/978-1-4419-6530-1_3.

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

Birenbaum, M. (1996). Assessment 2000: Towards a pluralistic approach to assessment. In M. Birenbaum & F. J. R. C. Docy (Eds.), *Alternatives in assessment of achievements, learning processes and prior knowledge* (pp. 3–30). Boston: Kluwer.

Blomeke, S., Gustafsson, J., & Shavelson, R. J. (2015). Beyond dichotomies: Competence viewed as a continuum. *Zeitschrift fur Psychologie, 223*(1), 3–13.

Brewer, L. (2013). *Enhancing the employability of disadvantaged youth: What? Why? and How? Guide to core work skills*. Geneva: International Labour Organization.

Buchner, A. (1995). Basic topics and approaches to the study of complex problem solving. In P. A. Frensch & J. Funke (Eds.), *Complex problem solving: The European perspective* (pp. 27–63). Hillsdale: Erlbaum.

Care, E., & Griffin, P. (2014). An approach to assessment of collaborative problem solving. *Research and Practice in Technology Enhanced Learning, 9*(3), 367–388.

Care, E., & Griffin, P. (2017). Collaborative problem solving processes. In B. Csapó & J. Funke (Eds.), *The nature of problem solving. Using research to inspire 21st century learning* (p. 227–243). Paris: OECD Publishing.

Care, E., Griffin, P., Scoular, C., Awwal, N., & Zoanetti, N. (2015). Collaborative problem solving tasks. In P. Griffin & E. Care (Eds.), *Assessment and teaching of 21st century skills: Methods and approach* (pp. 85–104). Dordrecht: Springer.

Care, E., & Luo, R. (2016). *Assessment of transversal competencies: Policy and practice in the Asia-Pacific Region*. Bangkok: UNESCO.

Care, E., Anderson, K., & Kim, H. (2016a). *Visualizing the breadth of skills movement across education systems. Skills for a changing world*. Washington, DC: The Brookings Institution.

Care, E., Scoular, C., & Griffin, P. (2016b). Assessment of collaborative problem solving in education environments. *Applied Measurement in Education, 29*(4), 250–264.

Catts, R., & Lau, J. (2008). *Towards information literacy indicators: Conceptual framework paper*. Paris: UNESCO.

Csapó, B., Lörincz, A., & Molnár, G. (2012). Technological issues for computer-based assessment. In P. Griffin & E. Care (Eds.), *Assessment and teaching of 21st century skills* (pp. 143–230). Dordrecht: Springer.

Darling-Hammond, L., & Snyder, J. (2000). Authentic assessment in teaching in context. *Teaching and Teacher Education, 16*, 523–545.

Dörner, D. (1986). Diagnostik der operativen Intelligenz [Diagnostics of operative intelligence]. *Diagnostica, 32*, 290–308.

Ercikan, K., & Oliveri, M. E. (2016). In search of validity evidence in support of the interpretation and use of assessments of complex constructs: Discussion of research on assessing 21st century skills. *Applied Measurement in Education, 29*(4), 310–318.

Fraillon, J., Ainley, J., Schulz, W., Friedman, T., & Gebhardt, E. (2014). *Preparing for life in a digital age: The IEA International Computer and Information Literacy Study international report*. Cham: Springer.

Fraillon, J., Schulz, W., Friedman, T., Ainley, J., & Gebhardt, E. (2015). *ICILS 2013 technical report*. Amsterdam: IEA.

Funke, J. (2001). Dynamic systems as tools for analysing human judgement. *Thinking & Reasoning, 7*, 69–89.

Funke, J. (2003). *Problemlösendes Denken [Problem solving and thinking]*. Stuttgart: Kohlhammer.

Gee, J. P. (2010). Human action and social groups as the natural home of assessment. In V. J. Shute & B. J. Becker (Eds.) (2010), *Innovative assessment for the 21st century* (pp. 13–40). Boston, MA: Springer Science+Business Media. doi:10.1007/978-1-4419-6530-1_3

Greiff, S., & Funke, J. (2009). Measuring complex problem solving: The MicroDYN approach. In F. Scheuermann & J. Björnsson (Eds.), *The transition to computer-based assessment: Lessons learned from large-scale surveys and implications for testing* (pp. 157–163). Luxembourg: Office for Official Publications of the European Communities.

Greiff, S., Wüstenberg, S., & Funke, J. (2012). Dynamic problem solving: A new assessment perspective. *Applied Psychological Measurement, 36*, 189–213.

Greiff, S., Wustenberg, S., Holt, D. V., Goldhammer, F., & Funke, J. (2013). Computer-based assessment of complex problem solving: Concept, implementation, and application. *Educational Technology Research and Development, 61*, 407–421.

Gulikers, J. T. M., Bastiaens, T. J., & Kirschner, P. A. (2004). A five-dimensional framework for authentic assessment. *Educational Technology Research and Development, 52*, 67–86.

Gulikers, J. T. M., Bastiaens, T. J., Kirschner, P. A., & Kester, L. (2006). Relations between student perceptions of assessment authenticity, study approaches and learning outcome. *Studies in Educational Evaluation, 32*, 381–400.

Hesse, F., Care, E., Buder, J., Sassenberg, K., & Griffin, P. (2015). A framework for teachable collaborative problem solving skills. In P. Griffin & E. Care (Eds.), *Assessment and teaching of 21st century skills: Methods and approach* (pp. 37–56). Dordrecht: Springer.

Kozma, R. B. (2011). A framework for ICT policies to transform education. In *Transforming education: the power of ICT policies* (pp. 19–36). Paris: UNESCO.

Lundberg, S. (2015). *Non-cognitive skills as human capital*. NBER/CRIW Conference on Education, Skills, and Technical Change.

Mayer, R. E., & Wittrock, M. C. (2006). Problem solving. In P. A. Alexander & P. H. Winne (Eds.), *Handbook of educational psychology* (pp. 287–303). Mahwah: Erlbaum.

OECD. (2013). *PISA 2015 draft collaborative problem solving framework*. Paris: OECD. https://www.oecd.org/pisa/pisaproducts/Draft%20PISA%202015%20Collaborative%20Problem%20Solving%20Framework%20.pdf

Parker, R., & Fraillon, J. (2016). *Southeast Asia Primary Learning Metrics (SEA-PLM): Global citizenship domain assessment framework*. Melbourne: Australian Council For Educational Research. http://research.acer.edu.au/ar_misc/20.

Pearson. (2017). *Edexcel International GCSE Global Citizenship (2017)*. Retrieved from http://qualifications.pearson.com/en/qualifications/edexcel-international-gcses-and-edexcel-certificates/international-gcse-global-citizenship-2017.html

Pellegrino, J. W., & Hilton, M. L. (2012). *Education for life and work: Developing transferable knowledge and skills in the 21st century*. Washington, DC: National Research Council, National Academies Press. http://nap.edu/13398.

Pellegrino, J. W., DiBello, L. V., & Goldman, S. R. (2016). A framework for conceptualizing and evaluating the validity of instructionally relevant assessments. *Educational Psychologist, 0*(0), 1–23.

Ramos, G., & Schleicher, A. (2016). *Global competency for an inclusive world*. Paris: OECD.

Rosen, Y. (2015). Computer-based assessment of collaborative problem solving: Exploring the feasibility of human-to-agent approach. *International Journal of Artificial Intelligence in Education, 25*, 380–406.

Rosen, Y., & Foltz, P. (2014). Assessing collaborative problem solving through automated technologies. *Research and Practice in Technology Enhanced Learning, 3*(9), 389–410.

Schwartz, H., Hamilton, L. S., Stecher, B. M., & Steele, J. L. (2011). *Expanded measures of school performance*. Santa Monica: RAND Corporation.

Scoular, C., Care, E., & Hesse, F. (2017). Designs for operationalizing collaborative problem solving for automated assessment. *Journal of Educational Measurement, 54*(1), 12–35.

Sluijsmans, D. M. A. (2002). S*tudent involvement in assessment. The training of peer assessment skills*. Unpublished doctoral dissertation Heerlen, The Netherlands: Open University of the Netherlands.

Soland, J., Hamilton, L. S., & Stecher, B. M. (2013). *Measuring 21st century competencies: Guidance for educators*. New York: RAND Corporation/Asia Society/Global Cities Education Network.

UNESCO. (2014). *Global citizenship education: Preparing learners for the challenges of the 21st century*. Paris: UNESCO.

von Davier, A. A., Zhu, M., & Kyllonen, P. (Eds.). (2017). *Innovative assessment of collaboration*. Cham: Springer International.

Wiggins, G. (1989). A true test: Toward more authentic and equitable assessment. *Phi Delta Kappan, 70*(9), 730–713.