NIE
NATIONAL INSTITUTE OF EDUCATION
SINGAPORE

Assessment of 21st Century Skills

Sima Aghazadeh

NIE Working Paper Series No. 14

About NIE Working Paper Series

The *NIE Working Paper Series* is intended as a means of regular communication between the mutually dependent spheres of theory and practice in education. Forward- and outward-looking, the Papers are conceptualized with a local issue at hand, and will survey international and local state of thought to assemble a principled response appropriate for our context. The intended audience for this publication are policymakers, school leaders and practitioners with an interest in how theoretical and empirical perspectives inform practice. The Working Papers are published in a highly readable style, and appended with an expanded exposition and comprehensive reference for readers who want to know more.

Suggested citation: Aghazadeh, S. (2019). *Assessment of 21st Century Skills* (NIE Working Paper Series No. 14). Singapore: National Institute of Education.

ISBN: 978-981-14-5320-5

Assessment of 21st Century Skills

Sima Aghazadeh

Abstract

In response to the rapid process of globalisation and knowledge-based economies alongside the transformational development of information and communication technologies (ICT) in our life, several frameworks have been developed to address competencies or skills required for success in 21st century society and workplaces. In line with such demands, learning, teaching, and assessment of 21st century skills have become urgent; however, assessment of 21st century skills is still in its infancy and one of the weakest points in current efforts to integrate these skills in the school curricula. Thus, assessment reform clearly needs more attention in policy and practice. To that end, this paper aims to focus on the potential and pragmatic assessments of 21st century skills.

This paper consists of 4 sections. Section 1 discusses the necessary characteristics for 21st century skills assessment which determine the types, criteria, and purpose of the tests. Section 2 summarises different assessment methods—self-ratings, others' ratings, portfolios, situational judgment tests, performance tasks, computer and game-based assessment—along with their advantages and limitations. However, given that today's young generation of learners are digital natives, growing up with computers, video games, and social media, it is important that this paper focuses on technology-based assessment. In this regard, Section 3 is dedicated to technology-based assessment, comprising computer-based and game-based methods. It presents examples of each in assessing 21st century skills. Lastly, in section 4, the paper proceeds to outline the major challenges that need to be addressed in employing different methods to assess 21st century skills. It also offers suggestions and implications in terms of policy, practice, and theory for researchers, educators, and policymakers who wish to assess 21st century skills in classroom.

Introduction

In response to the rapid process of globalisation and knowledge-based economies alongside the transformational development of information and communication technologies (ICT) in our life, work, and education, numerous organisations, scholars and educational communities have developed several frameworks to address competencies or skills required for the 21st century (e.g., Binkley, Erstad, Herman, Raizen, & Ripley, 2012; Partnership for 21st Century Skills (2009); Tan, Choo, Kang, & Liem, 2017). Each framework seeks to provide a list of skills or competencies[1] necessary to prepare young learners for a more technology-driven, interconnected, and competitive workplace. Appendix 1 shows the most popular and frequently cited 21st century frameworks and their classifications of skills.

21st century skills are framed and conceptualised with different focus, areas of emphasis, or nature of content. For example, some frameworks seek to define a set of skills based on student behaviours (e.g., creativity might refer to openness and responsiveness to new ideas), while others refer extensively to skills (e.g., creativity might refer to the ability to develop innovative and creative ideas), or some focus more on specific knowledge (e.g., creativity might be knowledge of a wide range of creation techniques) (Binkley et al., 2012). At times, different terminologies are used to refer to the same set of skills. Therefore, there is no single definitive list of 21st century skills. Nevertheless, a preliminary literature review and the findings of some studies that compared different frameworks (Voogt & Roblin, 2012; Dede 2010; Trier, 2003) reveal that there are common skills running through various lists which consider communication, collaboration, ICT (including information literacy, technological literacy, and ICT literacy), creativity, critical thinking and problem-solving as the essential 21st century skills (Griffin, Care, & McGaw, 2012; OECD 2005; Trilling & Fadel 2009). Of the key skills, some, like collaboration and communication, have been identified as skills that support the observed shifts in the way people work, especially through digital technologies (Binkley et al., 2012; O'Neil & Chuang, 2008). Research conducted and synthesised over the last decade confirm that ICT and

1. Athough there is a difference between the terms skill and competency, as one provided by the OECD's DeSeCo project (www.oecd.org), different frameworks use the terms interchangeably or with slightly different definitions. Some organisations prefer to use the term competency (e.g. Ministry of Education of Singapore, 2010; OECD, 2005; UNESCO, 2012), with the aim to refer to a bigger scope of knowledge, skills and attitudes. This paper will use both terms interchangeably, depending on the context of the research and case studies being discussed.

collaborative problem-solving are requisite for success in both school and work (O'Neil & Chuang, 2008).

In line with the demands of the 21st century society, learning, teaching and assessment of 21st century skills have become urgent. In their study, Ananiadou & Claro (2009) found that across 17 OECD (Organisation for Economic Co-operation and Development) countries, most of them had adopted 21st century competencies (CC) in their curriculum. Gordon et al. (2009) reported a similar conclusion when studying the implementation of 21st CC across the education systems of 27 Member States of the European Union. Similarly, Care & Luo (2016) demonstrated that 86% of the 102 countries represented by the Brookings Institution included 21st century skills in some aspects of their educational goals and designs. These studies demonstrate the ubiquitous presence and identification of 21st century skills in most countries with the global uptake of the notion of skills education. However, fewer countries have successfully integrated these skills in the curriculum and even fewer have changed approaches to assess these skills. A study supported by UNESCO through NEQMAP across the nine participating countries in the Asia Pacific region (Care & Luo, 2016) with focus on the implementation of assessment of 21st century skills, has reported that there is strong evidence of awareness of the need for assessment at policy and school levels, but many challenges—political, pedagogical and technical—hamper its implementation. On a related note, Comfort & Timms (in Care, Griffin, & McGaw, 2018) claim that despite large scale 21st century initiatives such as Partnerships 21 and ATC21S, teaching and assessment of 21st century skills in the U.S. is minimal.

It has been agreed that an integral part of developing and fostering 21st century skills is assessment, which should be used to support the process of day-to-day classroom teaching and learning. Despite this, assessment of 21st century skills is still in its infancy (Care & Kim, 2018) and one of the weakest points in current efforts to integrate 21st century skills in the school curricula (Ananiadou & Claro, 2009; Gordon et al., 2009). There is no assessment of 21st century skills in widespread use yet, and it is one of the major challenges in the implementation of these skills (Tan, Choo, Kang, & Liem, 2017). Nonetheless, the emerging progress is undeniable. Assessment reform

clearly needs more attention in policy and practice. Therefore, this paper aims to shift the focus beyond the theoretical and conceptual definition and enhancement of 21st century skills toward potential and pragmatic assessments of these skills, with the focus on common skills, categorised in Appendix 1. It presents different methods that educators can use to assess 21st CC through a literature review of international and local studies.

To accomplish its objective, this paper consists of 4 sections. Section 1 discusses the necessary characteristics for 21st century skills assessment which determine the types, criteria and purpose of the tests. Section 2 summarises different assessment methods—self-ratings, others' ratings, portfolios, situational judgment tests, performance tasks, computer and game-based assessment—along with their advantages and limitations. However, given that today's young generation of learners are *digital natives* (Prensky, 2001), growing up with computers, video games, and social media, it is important that this paper focuses on technology-based assessment. In this regard, Section 3 is dedicated to technology-based assessment, comprising computer-based and game-based methods. It presents examples of each in assessing 21st century skills. Lastly, in section 4, the paper proceeds to outline the major challenges that need to be addressed in employing different methods to assess 21st century skills. It also offers suggestions and implications in terms of policy, practice and theory for researchers, educators and policymakers who wish to assess 21st century skills in classroom.

Section 1: Transforming Assessment to Measure 21st Century Skills

The goal of 21st century education is to prepare students for the future workplace, thus, the assessment of the required skills must be designed in a way that can provide necessary information to determine students' ability to cope with real-life situations. To that end, the current and standardised practice of assessment (i.e., individual testing through memorising or recalling facts, or applying simple procedure) fails to meet this purpose. To ensure that students actually acquire and apply 21st century skills, we need a wide spectrum of assessments to evaluate their performance in the required areas, equip them with the ability to interact with peers and trained professionals, and provide

them with appropriate and timely feedback that is seamlessly integrated into the learning experience (Rupp, Gushta, Mislevy, & Shaffer 2010).

Over a dozen years ago, the National Research Council also called for changes in the assessment of 21st CC in which three things are suggested: 1) combine learning processes; 2) change the focus of assessment from *what is easily measured* to *what is most highly valued*; and 3) emphasise more on learners' engagement in ongoing assessment of their work and that of others[2]. Such changes are yet to be addressed adequately. To this end, this section seeks to: 1) summarise the main characteristics of such transformative assessment; and 2) elaborate three fundamental properties of assessment required for 21st century skills, as pointed out by Shaffer & Gee (2012).

1.1 Characteristics of 21st Century Skills Assessment

This section elaborates on the major characteristics and requirements of 21st century skills assessment, highlighted in the literature. (For the purpose of clarity, the key elements are marked in bold.)

As the largest internationally coordinated effort, ATC21S (sponsored by Cisco, Intel and Microsoft) has attempted to define new assessment approaches and the developmental learning progressions underpinning them. To achieve this aim, the project was planned to consist of five phases: 1) conceptualising of 21st century skills; 2) skill identification and hypotheses formation; 3) development and coding via cognitive laboratory[3] ; 4) pilot studies and trials; and 5) dissemination scale and policy (Griffin et al., 2012; Griffin & Care, 2015). The authors believed that the strategies and tasks used in this process should be "open-access, open-source, and prototype versions" as a shift towards **a more digital and technology-enhanced assessment practice** (Griffin et al. 2012, p. 22). The ATC21S project focuses on measuring two skills areas—collaborative problem-solving and ICT literacy in digital networks—by providing an approach in which students are engaged with and assessed through online and collaborative tasks. Students' digital and network literacy, and problem-solving skills and their **developmental progressions** (students' continuum progress from

2. National Research Council (NRC): https://www.nap.edu/catalog/13215/assessing-21st-century-skills-summary-of-a-workshop

3. The cognitive laboratory (coglabs) engaged students and teachers to complete the tasks with 'think aloud' and group discussion. The purpose of the cognitive laboratory was to ensure that assessment tasks had the capacity to elicit evidence and identify potential coding categories for automatic scoring and data retrieval (see more in Griffin et al. 2012)

lower to higher level of proficiency) are captured electronically in log files.

Similarly, Wilson, Scalise, & Gochyyev (2018), while providing contextual information about the ATC21S project, and its goals and framework for ICT literacy in networks, also described four principles of good assessment applicable to measuring 21st century skills: 1) to be based on a **developmental progression** which shows student developmental understanding of particular concepts and skills; 2) to be **aligned with pedagogical goals** which establishes a good match between what is taught and what is assessed; 3) to produce **valid and reliable results** of what students know and can do; and 4) to provide information useful to both teachers and students. The authors emphasised a **developmental paradigm**, or learning progression, in the assessment process as the theoretical maps of intended constructs in terms of depth, breadth and how the skills change as they mature for students. Teachers must have enough evidence to explain the results to students and make proper inferences. Students must participate in the assessment process which is designed to simultaneously encourage the development of targeted skills (Wilson et al., 2018). The authors also argued that there is a need for the ATC21S project to base its research on the definition of these learning progressions and identify where on a learning progression students are located. Wilson (2009) presented definitions and methods for developing learning progressions, as did a National Research Council report (NRC, 2007).

The ATC21S framework further explained that assessments of 21st century skills need to have the following general characteristics, as summarised by Voogt & Roblin (2012): (a) be aligned with the development of educational goals of the 21st century; (b) be adaptable, unpredictable and responsive to new developments; (c) be largely **performance based**; (d) provide **productive and usable feedback** for all intended users and contribute to teaching and learning progress; and (e) meet good assessments criteria (i.e., **be fair, technically sound, valid, and reliable**). In line with that, Greenstein (2012) encapsulated the fundamentals of assessment of 21st century skills as being responsive, flexible, integrated, informative and communicated, technically sound, systematic, and through **multiple methods** where students can demonstrate knowledge and skills through relevant tasks

and performances.

In their recent analysis, Lai & Viering (2012) also recommended the incorporation of **multiple measures** to assess 21st century skills so that it can "permit triangulation of inferences; designing complex and/or challenging tasks" (p. 1). They stressed on the **open-ended and ill-structured tasks** in **authentic contexts** in which students' thinking and reasoning will be made visible, and innovative and technological-enhanced approaches can be utilised. They argued that some of the 21st century skills are difficult to define and measure; any single assessment method has its limitations in their attempt to measure these skills (which will be discussed in Section 2). Research suggests that several of these skills (e.g., creativity, collaborative problem-solving, critical thinking) are complex and entail cognitive and non-cognitive and affective components. Due to the complexity of these skills, the use of multiple measures is recommended (Treffinger, Young, Selby, & Shepardson, 2002).

To summarise the characteristics and requirements of 21st century skills assessment and its tasks, assessment designs must be **authentic** and deal with real-world problems so that assessment tasks actually stimulate, prompt and facilitate the capture of the targeted skills (Care & Kim, 2018; Care et al., 2018). To achieve this, what and how we test must be integrated by tracking different kinds of information from different sources about students over time to assess their ability to apply relevant knowledge, skills and attitudes in real-life situations successfully.

Relatedly, assessment tasks must be **complex and challenging enough** to promote student engagement, motivation, as well as their cognitive and critical thinking skills (Lai & Viering, 2012). Providing complex tasks allow students to discuss and negotiate their learning standpoints and give reasoning or justification for their views (necessary in communication and collaboration). Under such terms, students will be able to see the multiple representations of concepts and identify or resolve the conceptual conflicts.

Validity and reliability are two important requirements of any effective assessment which enable us to make inferences about what learners

have learnt or how well they have learnt it (Griffin & Nix, 1991). Validity investigation in 21st century skills assessment is a central concept because it helps identify sources of construct-irrelevant variance, that is, instances in which results are influenced by a skill or attribute other than what is expected to be measured (Soland, Hamilton, & Stecher, 2013). While measuring a skill of interest, a test may measure some aspect of that competency consistently but fail to capture other essential aspects of that competency. Thus the question is whether the construct itself is being assessed, or merely some of its components, or to what degree the measured competency reflects all other aspects (Soland et al., 2013). For example, the measurement of collaborative problem-solving, both by PISA OECD and ATC21S, found the social components or subskills elusive while other aspects of the construct were captured. A test with low levels of reliability does not provide useful information about students' skills in the tested area. One way to reach an acceptable reliability is to use rubrics (holistic or analytic)[4] to score students' responses in assessing skills and their subskills (Saxton, Belanger, & Becker, 2012). Another way which can dramatically increase both reliability and validity is a multi-method approach to measurement (Lai & Viering, 2012; Eid & Diener, 2006) although it might be argued of not being time and cost effective.

The **use of technology** in assessment is regarded as another important feature to help combine learning, teaching and assessment as parts of one process. The National Research Council (NRC) workshop, held in 2009, delved more into the topic of 21st century skills assessment and explored assessment strategies. The Committee reviewed the assessment methods and related research for three clusters of skills (cognitive, interpersonal and intrapersonal) with special attention to recent developments in technology-enabled assessment of critical thinking and problem-solving skills (Koenig, 2011). Furthermore, both the P21 and the ATC21S frameworks emphasised the potential use of ICT and technology to support the assessment of 21st century skills, especially those complex skills (e.g., collaborative problem-solving or ICT literacy) which are hard to measure by traditional tests.

4. While a holistic rubric provides single descriptions for each performance with the focus on overall assessment, an analytic rubric is more granular as it assesses different skills and each criterion/dimension of the skill individually, then sums the individual scores to obtain a total score. The latter can provide a more detailed feedback, including students' strengths and weaknesses; but it is more time-consuming. In sum, a holistic scoring approach would be more suitable for summative assessment; analytic scoring rubrics for formative purposes. Although using analytic rubrics seems more desirable in skill assessment, it is noteworthy that one type of rubric is not inherently better than the other. Each can be used based on different implications like the purpose of assessment, time requirements, specific skill criteria, and so on. (more details available in Nitko, 2001; Montgomery, 2001).

The delivery of assessment can be more effective as ICT enables quicker results, reduces the cost and time required to score, and facilitates feedback. The ATCS framework suggests that technology can be used to improve assessment practices by changing the development, delivery and scoring of tests as well as the substance of assessments through more authentic tasks. From this perspective, ICT can be used to design a transformative assessment system to measure other 21st century skills.

The Costa Rica project, as a part of ATC21S, is a vivid example of using ICT as a tool to effectively use, develop, manage, and evaluate other competencies. A pilot project, trialled with a sample of 9,829 students in the 7th grade (13-year-olds), assessed their ICT, communication and problem-solving skills via an innovative and validated assessment platform. The results of this project showed the familiarity and exposure to ICT skills as one of the most important and difficult variables explaining the acquisition of other skills. Students with longer exposure to computer and IT programme (from 3 to 6 years) scored higher than students with less than 3 years of or no exposure at all in the three assessed domains. An online platform (www.fod.ac.cr/competencias21) was created to allocate the necessary resources, and formulated practical teaching and assessing guidelines with examples of how Costa Rican teachers were implementing them (Bujanda, Muñoz, & Zúñiga, 2018). This project has become a case study for the role of ICT in assessment reformation for many scholars and policymakers (for more details, see Bujanda, Muñoz, & Zúñiga, 2018).

1.2 Properties Required for 21st Century Skills Assessment

Shaffer & Gee (2012) argued that three fundamental properties of assessment needed for 21st CC are *what* is assessed, the *purpose* of assessment in the first place, and *how* the assessment takes place. The following will elaborate the first two properties, and section 2 will continue with the potential assessment methods.

1.2.1 What to assess: defining the skills

As mentioned earlier in the paper, the first step in the process of assessment of 21st century skills is to define the skills and their constructs. A report on the NRC workshop (Koenig, 2011) indicated

that giving specific operational definitions of constructs within the scope of 21st century skills is essential in the development of assessment tasks. According to the workshop committee, the definitions must be specific in clarifying whether the targeted skill or construct is domain-general or domain-specific. It means that it should be clear if students can solve problems or think critically in the presence or absence of that subject and transfer their skills from one context/domain to another (e.g., Kuncel, 2011). Moreover, the definitions and conception of some competences (e.g., communication or citizenship) have transformed into virtual communication or online citizens by the development of the internet and technologies (Loader, 2007).

Adequate and precise definition of skills determines the types of tasks to be used to elicit responses and performances (Wilson et al., 2012; Almond, Steinberg, & Mislevy, 2003). For assessment purposes, skills must be defined in measurable terms to design assessment tasks and determine the kind of information to be collected: "rigorous assessment methods cannot of course be developed without clear definitions of the skills and competencies in question" (Ananiadou & Claro, 2009, p. 16). Almond, Steinberg, & Mislevy, (2003) proposed the four-process architecture: 1) creating the tasks; 2) assigning values (codes or scores) to the student responses to these tasks; 3) gathering and delivering the responses; and 4) modelling and analysis of those responses, only after clarifying the skill's definition(s).

Different frameworks offer different conceptualisations of the skills which makes finding an all-encompassing definition and assessment more challenging. For example, collaboration is classified as a *learning* skill (P21, 2009), as an *interpersona*l skill (NRC in Koenig, 2011) or a *way of working* (ATC21S, 2009). These frameworks have different conceptualisations of collaboration as a construct, and its interaction with other skills (Lai & Viering, 2012). As an attempt to give clearer definitions of ATC21S skills, Binkley et al. (2012) have comprehensively set out the operational definitions of each skill, describing what they are and what relevant measurements for them might be. Several other attempts that have summarised the definition of 21st CC relevant to K–12 assessment include the following: the Collaborative for Academic, Social, and Emotional Learning (CASEL, 2019), the Chicago Schools consortium (for K–12) (Farrington et al., 2012), reviews conducted by

the National Research Council (e.g., Pellegrino & Hilton, 2012), OECD Programme for the International Assessment of Adult Competencies (PIAAC, 2018), and the Programme for International Student Assessment (PISA) (OECD, 2013).

1.2.1.1 Defining problem-solving skill: an example

One example to clarify the above-mentioned challenge is the necessity of having a precise definition of problem-solving skill and its different dimensions for assessment purposes. Problem-solving is a general term and has been defined with clear distinctions based on different dimensions and domains. Not all problems are the same; research differentiates between static and interactive/complex, or well-defined and ill-defined problem-solving (e.g., Funke, Fischer, & Holt, 2018). Another definitional dimension to be considered while measuring problem-solving skill is whether it is individual or collaborative (e.g., Griffin et al., 2012). Defining collaborative problem-solving is more challenging as the skill comprised problem-solving (cognitive), collaboration and consequently, communication (both social and affective) subskills in an interrelated or overlapping way. These cognitive and non-cognitive aspects are not easy to separate—conceptually or empirically (OECD, 2013). The next dimension to then consider is to know if the targeted collaborative problem-solving skill is a computer-based assessment and if so, whether it is in a human–human or human–agent (i.e., a simulated digital or virtual partner) setting (Rosen & Foltz, 2014; Rosen & Tager, 2013).

Rosen and Foltz (2014), and Rosen and Tager (2013) implemented one of the first empirical studies comparing human–human and human–agent collaborative problem-solving, and concluded that they are not equivalent. They observed different achievements even though the same tasks, communication methods and resources were employed in both settings. Based on their studies, in a human–agent setting, students achieved stronger performance results in discussions and disagreements as aspects of collaboration skill, but not necessarily in solving the problem. They concluded that human–human and human–agent settings can be useful for different assessment purposes—human–human for formative assessments, and human–agent for summative assessments as the computerised agents are designed to respond in more controlled and standardised ways.

The example of the challenge in defining problem-solving skill and its subskills in different contexts and domains prove that a detailed, precise and measurable definition of skills and their variables is vital in designing assessment tasks and methods. Educators should know what exactly they want to assess, what other skills/subskills might be involved, which aspect of it can or cannot be measured, and in what setting it could be used (this challenge will be further discussed in section 4).

1.2.2 Purpose of assessment: summative and formative

Summative assessments focus on making judgments about how well individuals do at the end of an instructional sequence, which is also considered *assessments of learning* (Ecclestone, 2010). On the other hand, formative assessments, also defined as *assessment for learning* (Bennett, 2011), emphasise evaluating learners' progress during the process of learning. The role of formative assessment can be reinforced and extended by *assessment as learning*, which encourages students to monitor their own learning by using self/peer-assessment and reflection to decide what they know and can do (Earl, 2012). The goal of *assessment as learning* is to guide students to be active, self-regulated and critical assessors in their process of learning.

The assessment of 21st century skills can be both summative and formative. NRC stresses on both while the ATC21S & P21 frameworks emphasise the need to move more towards formative assessments as a way to make students' learning more visible and provide feedback that can be useful for both teachers and students. PISA's assessment of high school students' ICT literacy through various activities is one example of summative assessment on a global scale. Existing published assessments of critical thinking, such as the California Critical Thinking Skills Test[5] , the Cornell Critical Thinking Tests[6] , and the Watson-Glaser Critical Thinking Appraisal[7], utilise multiple-choice items to assess the components of critical thinking skills for a summative purpose. P21's recommendation to migrate summative assessments from the rote memorisation to higher-order skills, like critical thinking, is a promising direction. West Virginia's revamp of its

5. For more information, see Facione, P.A. (1990). The *California Critical Thinking Skills Test--College Level. Technical Report #2. Factors Predictive of CT Skills*. Retrieved from https://eric.ed.gov/?id=ED327550

6. For more information, see Hasinger, E. (2019). *Cornell Critical Thinking Test Guide*. Retrieved from https://www.tests.com/Cornell-Critical-Thinking-Testing

summative assessments to incorporate higher-order thinking skills is a supporting example[8].

On the other hand, the ATC21S technology-based assessment of collaborative problem-solving (CPS) is an example of formative assessment. The focus of the project was more on drawing inferences on how students solved the problem than only determining the outcome of the test. The whole set of assessment tasks provided teachers with feedback to interpret students' capacity in CPS skill and subskills, such that a profile of each student's performance could be developed for formative instructional purposes (Griffin & Care, 2015).

To match the developmental progression characteristics of 21st century assessment, ongoing formative assessments are regarded as significant in developing those skills and better at monitoring them over time throughout the learning process, if conducted correctly and in a timely manner. Formative assessments can be employed for all 21st century skills, in all kinds of learning environments (Scardamalia, Bransford, Kozma, & Quellmalz, 2012). The regular use of classroom formative assessment can improve student achievement in skill-based learning (Shute, Hansen, & Almond, 2008). Principles of formative assessment cover many components of teacher–student and student–student interactions in the classroom, which can enhance collaboration and communication skills (Wiliam, 2007). Providing frequent and constructive feedback to learners has also been found to significantly improve learning where students require intervention in terms of 21st century skills (Shute, Hansen, & Almond, 2008). The goal of 21st century skills pedagogy is to actively engage students in their learning of skills which in turn motivates them to become independent learners and that is the desired outcome of effective formative assessments (Roskos & Neuman, 2012; Ecclestone, 2010). The development of new technologies has facilitated the effectiveness of formative assessment (Woolf, 2010). For example, various technological opportunities are being integrated into the assessment of problem-solving or collaboration, online or face-to-face, and providing timely and diagnostic feedback (Roskos & Neuman, 2012).

7. Retrieved from https://www.assessmentday.co.uk/watson-glaser-critical-thinking.htm

8. See www.p21.org/storage/documents/p21-stateimp_assessment.pdf

A way to ensure an effective formative assessment is for teachers to use rubrics and checklists to assess student mastery of 21st century skills. It is necessary to highlight that designing valid and reliable rubrics with clear developmental progression might not be easy for teachers as it requires expertise. There are numerous online examples which need to be assessed for quality before implementation. However, there are some rubrics provided on the P21 website for teachers to explore and probably modify to suit their needs. Catalina Foothills School District in Arizona provides a series of rubrics to assess students' skills such as critical thinking, productivity and self-direction in real time. Lawrence Township of Indiana currently uses rubrics to evaluate interactive communication and self-direction, and New Technology High School has implemented rubrics for evaluating peer collaboration and teamwork, work ethic and written communication (www.P21.org).

The proponents of formative assessments claim that summative assessments might not be suitable for assessing 21st century skills because the overemphasis on summative assessments has resulted in the lack of support in providing individualised and constructive feedback during the process of learning (Baker, Chung, & Delacruz, 2012; Ecclestone, 2010). However, the challenges of formative assessments should be taken into consideration as well. Some of these challenges are:

- Giving relevant feedback in learning tasks involving complex skills such as critical thinking and complex problem-solving is challenging.
- The perceived tension between formative and summative assessment as the latter is considered more accountable for student achievement in many countries.
- The sources of formative assessment are usually self- and peer-assessment or teacher's report where limitations affect validity and reliability (the limitations will be discussed further in section 2).
- Conducting formative assessment in large and multi-racial/cultural classroom is challenging and needs extra attention. In a large class, it is more time-consuming for teachers to plan tests and use data effectively. Therefore, there is an increased chance of inconsistency in marking and grading.

Diverse student backgrounds (e.g., race and gender) might affect the desired goals of formative assessment as students' engagement, responsiveness and interactions would vary accordingly.

- Technology facilitates formative assessments but it can cause some limitations such as cost, staff time, resources, accessibility issues and technical glitches.

It is necessary to understand that different kinds of assessments are for different purposes. Assessments used for summative purpose emphasises reliability and accountability, assessing a limited number of performances and contexts. On the other hand, formative purpose of assessment emphasises overall validity, assessing more performances in a wider range of contexts. Each type of assessment has a role to play in improving teaching and learning, and needs to be part of a total balanced and blended assessment system, providing a clearer and more complete picture of each learner (Fletcher, 2007; Shute, 2009). Similarly, the participants of the 2009 NRC workshop noted that in order to ensure students' progress in 21st century skills, both summative and formative assessments need to be available to evaluate their performance (Koenig, 2011).

The research and case studies referred to in this paper seek to explore the potential of any attempt to assess 21st century skills beyond the dichotomy of summative and formative, pursuant to a valid and reliable assessment process. Therefore, creating a balance between formative and summative assessments of student achievement is required to make diagnostic and comprehensive assessments of their competencies (Shute, 2009; Gee, 2004).

Section 2: Assessment Methods

A literature review based on Gipps (1994), Mislevy (1994) and CEDEFOP (2008), offers an operational definition of assessment as "the process of making inferences about an individuals' knowledge, skills, attitudes or other constructs using information from one or more methods such as tests, observations, interviews, projects or portfolios with reference to pre-defined criteria" (Pepper, 2013, p. 3). Accordingly, this section identifies the most common and potential assessment methods found in the existing literature to assess the scope and range

of K-12 students' 21st CC for summative or formative purposes. It is noteworthy that there are occasional areas of overlap in the methods used, partly because the targeted skills are inter-related in complex ways. One method might assess multiple competencies in different domains, or some of these methods can supplement each other in different studies.

The methods listed in this paper include the following: self-ratings, others' ratings, portfolios, situational judgment tests (SJT), performance tasks (also known as performance-based assessment), computer-based assessment (CBA) and game-based assessment (GBA). Technology has expanded the flexibility and accessibility of some of these methods, for example e-portfolios or online performance tasks, which makes it difficult to draw a clear line between technology and non-technology-based methods. However, it is important for educators to be aware that adopting and adapting plural measurement methods can yield better results tailored to suit specific classroom needs and meet the educational challenges in assessing 21st century skills (Greenstein, 2012). Table 1 summarises each of the mentioned methods with their advantages and limitations. There are some additional explanatory notes to clarify some of the points mentioned in Table 1. For further examples of 21st century skills being assessed by the mentioned methods, see Appendix 2.

Table 1. Summary of assessment methods for 21st century skills

Method	Description	Purpose of Assessment	Format	Advantages	Limitations	References
Self-rating	Learners to rate themselves	Mostly formative or combined with other tests for summative	• Questionnaires mostly in the form of a Likert rating scale; • Yes/No & short constructed response also possible (pen paper or online)	• Ease of administration; • Pragmatic; • Cost effective	• Highly subjective subject to response styles, reference group effects and social desirability bias; • Faking response; • Acquiescence response style; • Rating scales and language load should be modified and simplified for young learners	Lipnevich, MacCann, & Roberts, 2013); Gordon et al. (2009); Kyllonen, (2012)
Others' rating	Others such as peers, teachers or parents to rate learners	Mostly formative or combined with other tests for summative	• Questionnaires mostly in the form of a Likert rating scale; • Yes/No & short constructed response also possible (pen-paper or online)	• Less biased and more predictive of future outcomes than self-ratings	• Subjective; • Halo effects; • Rating scales and language load to be modified and simplified for young learners	Lipnevich et al. (2013); Gordon et al. (2009); Kyllonen, (2012)

Method	Description	Purpose of Assessment	Format	Advantages	Limitations	References
Portfolios	A series of information compiled over a period of time showcasing learner's learning progress, outcomes performances in real-life contexts	Formative and summative	• Paper-based in traditional portfolios; • Online, integrating audio-visual files and internet links in e-portfolios	• Holistic; • Dynamic; • Constructive; • Authentic; • Valuable for project-based learning; • Positive impact on students' performance	• Requires clear directions and guidelines for the selection of assessment entries and criteria which otherwise would result in student confusion and frustration; • Requires significant man-hour in scoring and providing feedback; • Lack of commonly agreed appraisal criteria or standards[9] • Risk of plagiarism	Pepper, (2011 & 2013); Chang & Tseng (2009); Simon & Forgette-Giroux (2000)
Situational Judgment Tests (SJTs)	A situation/ problem presented in words or videos to see how examinees respond to that situation	Mostly formative	• Multiple-choice or rating scale in paper-based SJTs • Online scenarios with audio-video, slideshows, graphic organisers in multimedia SJTs	• Blending the measurement of cognitive and non-cognitive skills; • Based on critical situations and job-like performances thus highly authentic; • High predictive validity • Minimised biasedness and fake responses, and less culturally affected than self or other ratings	• Scoring challenge as there are no objectively correct or clear-cut answers; • Less appropriate for younger learners with lesser meta-cognitive skills of judgment	Lipnevich et al. (2013); McDaniel et al. (2007); Whetzel & McDaniel (2009)
Performance tasks[10]	Authentic tasks such as exhibitions, experiments, group work, interviews, plays, presentations, projects and role plays in realistic contexts	Formative and summative	Face-to-face or online	• Social desirability, acquiescence bias, and faking; • More authentic, complex, and interactive: students to find resources, analyse and synthesise information to design/ create solutions to complex problems, resulting in higher critical thinking, problem-solving and soft skills	• Challenging on a wider scale or for high stakes assessments because of variation in teachers' judgements within and between schools and its negative impact on reliability; • Prerequisite skills & professional development necessary to teach and assess performance tasks in targeted domain/ competencies; • Easily influenced by external factors[11]	Pepper (2013); Kyllonen (2012); Darling-Hammond, Adamson, & Abedi (2010)

9. Lack of uniformity in portfolio assessment (Tisani, 2008) is due to different reasons: different assessors evaluate portfolios differently, students might perform and interact differently at different times, or different scoring rubrics may result in different products. Lack of uniformity in portfolio assessing poses a threat to both validity and reliability (Pellegrino & Hilton, 2012).

Computer-based assessment (CBA)	Use of computer and technology to deliver and score tests	Formative and summative	Various formats (multiple choice, objective question types, essay, short answer, etc.)	• Faster delivery and scoring; • More efficient administration; • Flexible test formats available; Provides immediate feedback; • Enhancement of validity and reliability through the capture of process data • Facilitation of and the capture of hard-to-measure skills like collaboration, problem-solving, etc.	• Aligning schools' technology infrastructure to support CBA can be costly; • Negative impact of poor instructional technology design, and technological glitches; • CT literacy for students as prerequisite & different level ICT literacy affecting their assessment results; • Higher risk of plagiarism	Quellmalz & Pellegrino (2009); Wilson et al. (2012); Wilson & Scalise (2015); Ramalingam & Adams (2018)
Game-based assessment (GBA)	Use of video and digital game designs to measure performance	Formative and summative	Digital games	• Integrates learning and assessing through performance; • Provides immediate feedback, supporting formative assessment; • Direct focus on 21st CC like problem-solving, collaboration & interaction, communication (virtually), and ICT skills; • Combines learning & entertainment, triggering student perseverance and engagement which can reduce assessment stress[12] ; • Equitable as game resources, rewards, and penalties are offered equally to players, regardless of their backgrounds; • Opportunity to learn from failure	• Time-consuming and costly to design, develop, and implement; • Difficult to design engaging games while satisfying assessment requirements (e.g., fairness, validity, and reliability); • Challenging to assess the appropriate knowledge, skills, or abilities in the process of gaming due to open-ended nature of games; • Possible increase of violence, aggression, inactivity, and obesity while decreasing prosocial behaviours (Gentile et al., 2009)	Ke, & Wang (2017); Shaffer & Gee (2012)

The following points provide some explanatory notes to clarify some items in *Table 1*:

1. Self/others rating limitations

Faking response and tendency toward positive response bias, known as *acquiescence response style*, are the most common limitations of self- and others-rating, especially among younger learners as they are more inclined to be optimistic (Turner, 1995), or may not understand the questions and select answers out of pressure resulting in socially

10. One promising example of performance tasks and portfolio frameworks 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: http://www.performanceassessmentresourcebank.org/bin/performance-tasks.

11. Task performance may be influenced by other competencies (e.g., communication or hand-eye coordination), aspects of the environment in which the task is performed, or by the physiological state (e.g., time of day, classroom noise, hunger, etc.). On a related note, whether or not performance assessment can capture the true nature of the targeted skill in varying degrees of competency should be addressed (Care & Kim, 2018). In assessing multi-dimensional skills, it is necessary to know whether the skill itself is being assessed, or only some of its subskills. Subsequently knowledge, skills, attitudes and values which contribute to those skills must be taken into consideration in performance assessment (Ramos & Schleicher, 2016).

12. This advantage is associated with *flow theory*, as a natural foundation for motivation in games and learning (Csikszentmihalyi, 1990). The flow experience occurs when game players can be so engaged in goal-directed activities that they might lose time and be driven by joy and pleasure rather than an external reward. The relationship between flow and GBA is further explained in the additional explanatory notes, at the end of this section.

desirable responses (Duckworth & Yeager, 2015). He and Van De Vijver (2015) provided a helpful review of response styles (extreme and midpoint responses, and socially desirable responses) and their effects on self-reports among 76,887 participants from 18 countries. Faking response occurs as students incline more towards the best response like "strongly agree" with any statement that reflects a value. Those who almost agree with more acceptable attitudes or behaviours are assumed to be faking their answers based on conformity to group and social expectations (Lipnevich et al. 2013).

Researchers suggest several methods for reducing fake response. These include giving real-time warnings (Sackett, 2006), using a forced-choice format (Stark, Chernyshenko, & Drasgow, 2005), the Bayesian Truth Serum (BTS)[13] (Prelec, 2004), and anchoring vignettes (King & Wand, 2007). However, evidence for the effectiveness of these methods in controlling faking remains ambiguous and more research is required to provide evidence of their validity (Heggestad, Morrison, Reeve, & McCloy 2006). Besides that, these methods can be used only for limited aspects of competencies. For example, in assessing critical thinking or creativity, they might not reveal individuals' underlying reasoning for choosing a particular answer (Ku, 2009).

Another limitation in others' rating is the halo effect where parents or teachers rate along a number of dimensions or occasions, and their opinion in one area is affected by another (Zhuang, MacCann, Wang, Liu, & Roberts, 2008; Nisbett & Wilson, 1977). This limitation can be addressed when the same candidates are rated by multiple raters or individual raters to accumulate ratings for many candidates over time, which makes it more time-consuming and costly (Kyllonen, 2012).

2. Technology-supported portfolios, SJTs, and performance tasks

As mentioned earlier, technology has enhanced and facilitated assessment in different ways. E-portfolios, digital SJTs and online performance assessments are good examples of this claim. The following offers some examples of each, with their advantages and

13. BTS is a scoring system for eliciting and evaluating subjective opinions from a group of respondents answering multiple-choice questions by providing truth-telling incentives. It requires respondents' answers as well as their estimates of other respondents' answers to the same questions. The scoring formula then assigns high scores to answers whose actual frequency is greater than their predicted frequency (Prelec, 2004).

limitations found in relevant literature:

The *eSCAPE* project is an example of e-portfolio assessment in which a 6-hour collaborative design workshop replaced school examinations in design and technology for students aged 16 in 11 participating schools across England. Students worked individually but within a group context to create e-portfolios of their design proposals. The assessment evidence was captured via video, photo, voice, sketchpad and keyboard. During the 6 hours, each student developed their design prototype, and the Personal Digital Assistant (PDA) provided a record of their progress, interactions and self-reflections. The project resulted in 250 e-portfolios and the reliability of the assessment method was reported to be very high (Binkley et al., 2012). Another example of an online portfolio assessment is reported by National Research Council (2011), which was designed for high school students and used by the Envision High School in Oakland, California to assess critical thinking, collaboration, communication and creativity. All students are required to assemble a portfolio in order to graduate. In her study, Barbera (2009) interconnected students' e-portfolios in netfolio so students can assess their peers' works. Through a chain of co-evaluators created in this method, mutual and progressive improvement process was observed. Similarly, Garrett, Thoms, Alrushiedat and Ryan (2009) linked university students' portfolios in a social network, encouraging them to share and evaluate their works, which increased student motivation and performance.

Multimedia Emotion Management Assessment (MEMA) was developed as a multimedia adaptation of the STEM (science, technology, engineering, and mathematics). Scenarios and scripts were created and revised by an expert panel consisting of the STEM authors and assessment development staff (MacCann, Lievens, Libbrecht, & Roberts, 2016). Recent studies demonstrated that multimedia SJTs show stronger validity than written SJTs, especially for assessing interpersonal and affective skills (e.g., Christian, Edwards, & Bradley, 2010) as they can represent a richer medium including more social—verbal and nonverbal—cues, and paralinguistic information (Lipnevich et al., 2013). Moreover, use of technology enhances the degree of authenticity and originality of real-life environment. Multimedia scenarios can minimise the problem of construct-irrelevant variance in

text-based SJTs, which is caused by the impact of examinees' verbal, reading, and writing proficiencies on their performance. However, the influence of IT literacy on student performance cannot be overlooked in multimedia SJTs (Zhuang et al., 2008), and productions, equipment and administration of multimedia is relatively high in cost.

Aesaert, Van Niljen, Vanderlinde, and van Braak (2014) conducted a study in which they outlined the development of a performance-based digital test to measure ICT competency in 7th grade primary-school students in Flanders, Belgium. Another example of online performance tasks is the one developed by the International Society for Technology in Education (ISTE) (http://www.iste.org/standards.aspx), which assess students', teachers' and administrators' level of skill in using ICT. In Hong Kong, a study was conducted by using online performance assessment of students' information literacy skills as part of the evaluation of the effectiveness of its IT in education strategies (Law, Yuen, Shum, & Lee, 2007).

3. Example of performance-based assessment

Singapore is an example of a national effort to move towards authentic performance assessment following the education policy by the Singapore Ministry of Education's *Thinking Schools, Learning Nation* in 1997 which aims to cultivate creative and critical thinking skills for the 21st century economy (Ng, 2004; Sharpe & Gopinathan, 2002). Research shows that the change of assessment practice towards performance tasks has been vital in generating and developing students' critical thinking and problem-solving skills as they are tested on their skill application in real-life situations (Koh, Tan, & Ng, 2012; Darling-Hammond et al., 2010). When designed and practised well, performance-based assessments reflect more intellectually on learning goals and include more authentic and open-ended tasks such as "sustained written prose where students are asked to elaborate on their understanding, explanations, arguments, and/or conclusions" (Koh et al., 2012, p. 140). In their report on implementing new assessment strategies in primary and secondary schools in Singapore, Fan et al. (2008) declared that the quantitative and qualitative data indicated that the performance tasks integrated in the mathematics classroom (with 160 Secondary 1 students and 4 teachers involved) were helpful in developing students' skills in higher-order thinking, problem-solving and

self-reflection in learning. Moreover, both the teachers and students had positive views about the value and feasibility of integrating performance assessment into their daily class activities.

4. GBA & Flow theory

Research has indicated that some common game elements, such as challenge, control, and fantasy influence players' motivation (e.g., Malone & Lepper, 1987), and motivated players are engaged enough to experience the state of "flow" (Csikszentmihalyi, 1990). According to Csikszentmihalyi (1990), the major components of the theory of flow consist of the following: a challenging activity requiring skill and concentration; clearly-defined goals of activities; immediate feedback; a sense of control; as well as loss of self-consciousness and time (i.e., immersion). Similarly, the positive experience of flow associated with digital games can induce motivation, engagement and consequently an optimal learning state promoted by clear goals, appropriate levels of challenge, and direct and consistent feedback (Shute et al., 2017). The idea is to exploit the characteristics of game environments to create and maintain the flow needed to keep the students engaged in solving more complex tasks. In other words, flow in GBA can be used to facilitate student knowledge as well as skill acquisition through their joy of learning.

Many educational games might interrupt the state of flow by inserting quizzes and tests (Shute, 2009). However, when assessments are conducted unobtrusively within educational games (i.e., stealth assessment that will be discussed in section 3), learning via gameplay can continue fluidly and flow is maintained (Shute & Ventura, 2013). Such assessments are intended to support learning and reduce test anxiety while not sacrificing validity and reliability (Shute, Ventura, Bauer, & Zapata-Rivera, 2009). In fact, incorporating the concept of flow in computer games as a model for evaluating player enjoyment has been a focus of recent studies (Cowley, Charles, Black, & Hickey, 2008; Sweetser & Wyeth, 2005). Sweetser and Wyeth (2005) considered that the concept of player enjoyment is similar to flow, therefore enjoyment (or flow) can be used as a model to design and evaluate games. They called the model *GameFlow*. In a similar way, Inal & Cagiltay (2007) referenced *GameFlow* to explain how to facilitate flow experiences in computer games. Following the *GameFlow* model, Fu, Su, and

Yu (2009) developed *EGameFlow*, a scale specifically to measure a learner's enjoyment of e-learning games. *EGameFlow* contains 42 items allocated into Csikszenmihalyi's flow components. The result of their study in a university's online learning course showed that the validity and reliability of the *EGameFlow* scale were satisfactory.

In short, if assessment is well-designed within the game environment, it will be able to facilitate the acquisition of knowledge and skill through the flow state which increases joy of learning in students.

Section 3: Technology-based Assessment

Today's generation is growing up with laptops, tablets, cell phones, and video games, and increasingly uses them in their daily interactions. As such, the pervasive presence of technology and multimedia has paved the way for technology-based learning and assessment in education. It has opened up new possibilities to assess 21st century skills in assessment functions, such as delivery and scoring, as well as in the measurement of the application of these skills. In the late 1990s, researchers began investigating on measuring complex problem-solving and higher order thinking skills by using technology (Baker et al., 2012; Quellmalz & Pellegrino, 2009). Based on current needs, more advanced and innovative assessments of competencies are expected to include different ICT technologies as integral components of the process (Wilson et al., 2018).

As shown in the previous section, some assessment methods can be carried out more efficiently by using technology such as e-portfolios, multimedia SJTs or online performance tests. There are also examples of web-based peer assessment strategies (Tsai, Lin, & Yuan, 2001). Technology has transformed the existing assessment methods, offered new tools to reflect and measure skills that have either been hard to assess or emerged as part of the information age such as ICT literacy and provided more authentic assessment tasks (Quellmalz & Pellegrino, 2009). The use of simulation and visualisation tools such as audio, video, animation and games, if developed properly, create more dynamic and real-life situations which cannot be achieved in traditional tests. Therefore, they can provide a wide range of rich interactions with the capability to assess more complex or more realistic situations.

This section presents technology-based assessment for 21st century skills either in the form of computer-based assessment (CBA) or in the more innovative form of game-based assessment (GBA). It offers some examples of CBA in the literature, and then focuses on GBA, features of a good game for assessment, examples of GBA for different skills and assessment purposes, and different models of games with examples (see Appendix 3 which offers some examples of research studies on GBA, coupled with their implications and limitations). However, before discussing the two forms of technology-based assessment, the following gives a brief summary of the innovative application of such assessment in different parts of the world.

In Asia, countries such as Singapore, Japan and Hong Kong, have been adjusting their curriculum and pedagogy in alignment with dynamic economic and global changes. Through focusing on innovation in learning process and assessment practice, these countries aim to prepare students for the knowledge economy (Plomp, Anderson, Law, & Quale, 2009). Singapore's *Thinking Schools, Learning Nation* (1997) and its first IT Masterplan in education, the Hong Kong SAR government launching a comprehensive curriculum reform in 2000 (EMB 2001), and Japan's e-learning strategy are vivid examples (Plomp et al., 2009). There have also been some small-scale explorations by researchers on technology-based assessment. One instance is the 2007 Hong Kong project on performance assessment of students' IT literacy as part of the IT evaluation in education strategy in Hong Kong (Law, Yuen, Shum, & Lee, 2007; http://il.cite.hku.hk/index.php). The participants in this assessment were Primary 5 (P5) and Secondary 2 (S2) students in the 2006/2007 academic year. The assessments administered to P5 students were a generic technical literacy assessment and IL (information literacy) in Chinese language and mathematics. For S2, they were a generic technical literacy assessment and IL in Chinese language and science.

The use of technology to assess knowledge and skills is growing in several European countries and the research and development unit of the European Union is facilitating these attempts (Scheuermann & Björnsson, 2009). In the 1990s, the National Institute for Educational Measurement in the Netherlands developed a student monitoring system through a software programme for primary education to assess

students' learning progress. Subsequently and as a consequence of national high-stake ICT-based testing, computer-based monitoring and assessment systems were developed and offered to public schools (Vlug, 1997).

Luxembourg was a pioneer in introducing a nationwide assessment system, moving immediately to online testing, skipping the paper-based assessment. The current version of the system is able to assess an entire cohort simultaneously. It includes an advanced statistical analysis unit and the automatic generation of feedback to teachers (Plichart, Jadoul, Vandenabeele, & Latour, 2004) called TAO (the acronym for Testing Assisté par Ordinateur, equivalent to Computer-Based Testing). This system has also been used in several international assessment programs, including the Electronic Reading Assessment (ERA) in PISA 2009 and the OECD Program for International Assessment of Adult Competencies (PIAAC) (OECD, 2013).

In Hungary, the first major technology-based testing took place in 2008 (Csapó, Molnár, & Tóth, 2009). An inductive reasoning test was administered to a large sample of 7th grade students in both paper-and-pencil and online versions (using the TAO platform). In 2009, a large-scale project was launched to develop an online diagnostic assessment system for the first six grades of primary school in reading, mathematics and science. The project included developing assessment frameworks, building assessment task banks and using technologies for migrating items from paper to computer (Csapó et al., 2009).

In the US, there are many examples in which technology is being used in large-scale summative testing for the primary and secondary levels. For example, The Measures of Academic Progress (MAP) (Northwest Evaluation Association), is a computer-adaptive test in reading, mathematics, language, and science (Van Horn, 2003). MAP is used by thousands of school districts. The test is linked to a diagnostic framework, DesCartes, which anchors the MAP score scale in skill descriptions that appear to offer formative information. Online tests offered by the major test publishers, mostly for formative assessments, include Acuity (CTB/McGraw-Hill), the PASeries (Pearson), and the Cognitive Tutors (Carnegie Learning). At the post-secondary level, the Graduate Record Examinations (GRE) General Test (ETS), the

Graduate Management Admission Test (GMAT, GMAC) and the Test of English as a Foreign Language (TOEFL) iBT (ETS) are all examples of computer-based and high-stakes tests used for educational purposes.

3.1. *Computer-based Assessment (CBA)*

Given the easier and faster assessment delivery and data collection in CBA[14] , migration from paper-and-pencil to the computer environment was an increasingly viable option (Scalise & Gifford, 2006; Wilson et al., 2012). Some prominent examples along these lines are the OECD PISA programs of Digital Reading assessments (OECD, 2013), Collaborative Problem-Solving assessments (OECD, 2013), and the U.S. National Assessment of Educational Progress technology-based pilot administrations in mathematics, reading, and science. In 2006, PISA began to use ICT in its assessment process in the domain of scientific literacy entitled a *Computer-Based Assessment of Science* (CBAS). The delivery of CBAS was administrated on laptop computers in schools, with the assessment system installed on a wireless or cabled network and one of the computers acting as the administrator's console. Student responses were saved both on the student's and the test administrator's computers. An online translation management system was developed to manage the translation and verification process for CBAS items (Haldane, 2009). Other examples of CBA platforms that allow educators to develop and deliver tests and quizzes, particularly for formative purposes, include *Questionmark Perception* (Bellotti et al., 2013), *ASSISTment* (Feng, Hefferman, & Beck, 2009), and *ACED* (Adaptive Content with Evidence-based Diagnosis) (Shute et al., 2008).

Table 2 summarises some examples of CBA from various literature as an attempt to assess 21st century skills. The literature shows that CBA has been employed for more than two decades and evolved from "automated administration or scoring of conventional test" to "integrated" and "transformative" assessment (Redecker, 2013, p. 3). The examples selected here belong to the latter and seek to explore

14. In this paper, CBA refers to both the method and mode of assessment, depending on what skills or what components of skill of interest are being measured, or determining how effectively students perform in a computerised environment. For example, when it is used for formative assessment or to assess certain constructs (e.g. collaborative problem-solving), it is more than a mode of delivery as it is embedded within or even central to the context and process of assessment and it allows test developers to do what is not possible with the pencil-and-paper version.

CBA for more complex and authentic problem contexts (most of them with the focus on formative assessment of ICT or IL skills) to support the assessment of other core skills.

Table 2: CBA examples to assess 21st century skills.

Example of CBA project	**Assessment description**	**Skill**	**Reference**
The *eVIVA*, developed at Ultralab in the UK	Formative assessment tools promoting self- and peer-assessment & self-regulated learning; students use the eVIVA website to set up an individual profile, take part in self-assessment activity, upload files and add comments to each other's work, create and record a milestone of what they have learnt. Teacher assessment is based on students' recorded milestones, e-portfolios, reflection on their and the other's works, any written answers attached to the questions, and classroom observations.	ICT literacy	Walton (2005)
A project commissioned by the Joint Information Systems Committee (JISC), as early as 2000, employed by the UK department of education	Authentic tasks assigned to students to complete after logging into the virtual desktop environment. For example, in one task students were asked to create a job vacancies page for the local virtual newspaper. Students had to research job vacancies across different websites within the virtual environment, collect information, send out virtual e-mails to clarify and confirm details, and respond to e-mails from the newspaper's virtual editor.	ICT and problem-solving	https://www.jisc.ac.uk/
Educational Testing service iSkills assessment (iSkills)	The *iSkills* assessment is delivered online in a secured testing environment. It presents scenario-based performance tasks in which students solve information problems using emails, web browsers, or presentation software. The assessment of ICT literacy focuses on different dimensions that are measured in a simulated software environment, including defining, accessing, managing, integrating, evaluating, creating and communicating using digital tools, communications tools, and/or networks.	ICT literacy	Katz & Macklin (2007)
Assess By Computer (ABC)	Using different question formats, it offers the opportunity to design a test via an interactive user interface where students can take the test on a stand-alone computer or within a web browser, encouraging self-regulated learning. It is designed to deliver and stimulate feedback for formative purposes, to support instructors to create, administer, assess, and analyse tests. For example, in a case study at the University of Manchester, ABC tool was used to assess international students' entry English communication. The data provided formative feedback both for learners to identify their weak areas of communication and teachers to shape their teaching through collaborative assessment.	Communication & collaboration	Bellotti et al. (2013) Wood (2009)
ICILS	Designed to measure international differences in students' computer and information literacy (CIL). This type of literacy refers to students' ability to use computers to investigate, create, and communicate in order to participate effectively at home, school, or in the workplace. Two key attributes are addressed: the test contents reflect real-world use of ICT, and the tests make use of the dynamic functionality and multimodal opportunities afforded by the computer-based environment.	ICT literacy	Fraillon, Schulz, Friedman, Ainley, & Gebhardt (2015)

The Student Tool for Technology Literacy (ST2L)	A performance-based assessment designed for middle school students; the assessment tasks include performance and some selected-response items, like text-based multiple-choice, true/false, and multiple-choice items with graphics and image map selections. Students respond to performance tasks in a software environment (e.g., Spreadsheet) that simulates real world application of ICT literacy skills. The tool is reported to be more suitable for low-stakes assessment but compared to paper-pencil or online self-report, provides a more objective and authentic measurement[15].	ICT literacy	Hohlfeld, Ritzhaupt, & Barron (2010)
World Class Tests	Commissioned by England's Department for Education & Skills (DfES), WCT are intended to assess problem-solving in the domains of mathematics, science, and design and technology for worldwide application. They have been adapted for children aged 8–14, and are now sold commercially under license in East Asia.	Problem-solving, critical thinking	http://www.worldclassarena.org Pead (2012)
The Tool for Real-time Assessment of Information Literacy Skills (TRAILS)	Conducted with 199 primary-five students (aged 10–11) from four schools in Hong Kong; all questions were close-ended, with two to four options each. Students' responses were collected through SurveyMonkey, an online survey tool administered by students' IT teacher during regular class hours.	Information literacy	Chu (2012)
Wiki-based writing assessment	Conducted with 25 secondary one students (aged 12–13) in Hong Kong; students' comments (text, embed videos, photos and quotes) and the frequency of their contribution on wiki were recorded, retrieved, and analysed for the purpose of assessment. Teachers could track changes made by each student by using 'history review' and 'version comparison'. This would help them grade students' performances objectively and also identify low achievers who needed more guidance.	Collaboration (patterns of activities in their inquiry-based project, their level and frequency of participation, as well as the degree of collaboration)	Chu, Yeung, & Chu (2012)
The PISA CPS test (2015)	The test measured collaboration, problem-solving, and how these two interact with each other to generate a desired outcome. Each student received multiple questions designed to measure the targeted competencies, and the assessment was intended for summative purposes. Each student was assigned a two-hour test form, an hour of which was devoted to CPS. Within CPS, units range from five-to twenty-minute collaborative interactions around a particular problem. For each unit, multiple measurements of communications, actions, products, and responses to probes were recorded. Each of these individual questions provided a score for one of the three CPS competencies.	Collaborative problem-solving	OECD (2013)

15. *The ST2L* has been used by more than 100,000 middle grade students in the state of Florida since its formal production release (Holmes, 2012; Hohlfeld, et al., 2010).

A pilot project of ATC21S, using The Berkeley Evaluation and Assessment Research (BEAR) web-based assessment	Two of the BEAR scenarios were selected: the science/math *Arctic Trek* to assess collaboration, and the literature/poetry analysis *Webspiration* to assess collaboration and digital literacy. In the *Arctic Trek*, students worked in team and studied on tools and approaches to unravel clues through the Go North/Polar Husky information website (www.polarhusky.com), having scientific and mathematics expeditions. The website focuses on space to represent itself and combines texting tools, chat, and dialogue for communication and collaboration. A question was given to students in teams. Roles would be assigned to each member and all the findings would be recorded in a Team Notebook (responses, clues, and explanations provided by students and any hints or assistance provided by teachers). Members used the web resources listed for them to answer the questions. Similarly, in the second scenario, *the Webspiration* online tool was used to assess collaboration while students formulated their own ideas on a given poem and created an idea map-collaboratively using the tool. Based on the preliminary results of these two scenarios, it was concluded that it was reliable and valid to measure collaborative digital literacy by using these two web-based methods[16].	Collaboration, Digital literacy	Wilson & Scalise (2015)

3.2. *Game-based Assessment*

Game-based assessment (GBA), if well-designed and well-developed, possesses the features that any *Good Assessment for Twenty-first century Education* (GATE) would have to possess (Gee, 2009; Shaffer & Gee, 2012). In the context of GBA, it is necessary to highlight two points. The first point to consider is whether to use existing games in the market or to design and customise a new game (e.g., *Urban Science* designed by Shaffer & Gee, 2012). For the first option, more research and development is needed before an existing game could be used as an assessment instrument. For example, investment in securing the reliability and validity of the measures used in the game is inevitable. Finding an existing game which can meet all assessment requirements is a big challenge. On the other hand, designing and developing a new game system that can serve for an assessment framework might be an ideal option, but it is costlier.

Second, GBA uses games for assessment purposes in two ways: either as assessment tools in themselves or embed them within the process of assessment by using other external methods. The former refers to the game's in-built features, such as scoring or levelling-up, that could be used to assess progress. The latter refers to integrating games with a variety of assessment methods (e.g., using a portfolio after playing the game or a questionnaire as a pre/post-test). Interviews

16. The research also emphasised that the elaboration of the learning progression of the targeted skills as well as empirical-based results will be calibrated when larger data sets are available. Then, the ATC21S resources will be available in the public domain to download, modify, and extend existing research (Wilson & Scalise, 2015).

(Chin, Dukes, & Gamson, 2009), knowledge maps (O'Neil, Chuang, & Chung, 2003), causal diagrams (Spector & Koszalka, 2004), multiple-choice questions and essays are the most common external assessment to which GBA can be integrated. Chen and Michael (2005) discussed three methods of integrating assessment into a serious game: completion assessment (when the player completed the lesson), process assessment (how the player chose actions or changed their minds) and teacher evaluation (based on teacher's observation of the student). However, their study did not offer any empirical evidence of such assessment methods.

When assessment is embedded, it becomes a part of the gameplay (compared to the external assessment which is not part of the game environment); thus, it does not interrupt the game and maintains the flow. Data on the learners' behaviours while playing the game can be recorded by clickstreams or log files (Chung & Baker, 2003), or information trail technique, a series of event markers deposited within any game at certain intervals over a period of time (Loh, Anantachai, Byun, & Lenox, 2007). Shute et al. (2009) stated that the term *embedded* refers to formative assessment to get accurate information about the learner, inserted into the game or curriculum in an unobtrusive way based on which students and teachers can act.

To provide an overview of major aspects of GBA in assessing 21st century skills, this section is divided into three subsections: GBA & 21st century skills, GBA and assessment purposes, and game models (stealth & epistemic models) for assessment. For readers interested to know more about GBA in practice, Appendix 3 provides some research studies of GBA, with the summary of their findings, implications, and limitations. However before going into the first subsection on GBA & 21st century skills, a brief discussion on serious games and the difference between the concept of GBA and gamification is in order.

Serious Games. Serious games are becoming more and more popular in different areas such as defense, education, scientific exploration, health care, and engineering. Although there is no single definition of the concept, it generally (at least in education) refers to the use of computer games to engage players for a specific purpose such as learning new knowledge and skills above entertainment (Susi,

Johannesson, & Backlund, 2007). It is noteworthy that while serious games refer to learning via play, GBA is the assessment of learning that can be inside the games or via external tools, as explained above. Gee (2009) outlined the six key properties for serious games necessary to create the motivational virtues of games in learning process as: "an underlying rule system and game goal to which the player is emotionally attached; micro-control that creates a sense of intimacy or a feeling of power; experiences that offer good learning opportunities; a match between affordance (allowing for a certain action to occur) and effectivity (the ability of a player to carry out such an action); modelling to make learning from experience more general and abstract, and encouragement to players to enact their own unique trajectory through the game" (p. 78).

In spite of their different perspectives towards serious games[17] , Shute et al. (2017) and Shaffer & Gee (2012), believed that a serious game must encompass interactive problem-solving, specific rules, and adaptive and sequential challenges (with a good mixture of practice and guidance), control, ongoing feedback, uncertainty and suspense, and sensory stimuli (audio-video, graphics, and storyline to excite the senses). These features position students/players to constantly work at the cutting (and exciting) edge of their abilities, and provide a good example of assessment as learning i.e. simultaneously learning and assessing experiences.

Bente and Breuer (2009) provided a detailed survey and analysis of serious games, their components and the related design techniques. After offering a literature review comparing different definitions and components of serious games and their relatives (e.g., e-learning, edutainment), they reasserted that the ideal serious game combines entertainment and learning such that players do not experience the learning as something external to the game. Therefore, implicit psychological situations and sensation, which are hidden or hard to aggregate can be made explicit in the process of gaming and taken into account for assessment purposes. The list of players' psychological measures includes arousal states, attention, workload, moods,

17. Shaffer & Gee consider games serious where they model professional practices and players think and act like real world professionals. They call it epistemic games. Shute et al. (2009) broaden the concept of serious games by considering entertainment games potentially useful for educational purposes. They outline two key features of serious games as being educational and immersive. They focus on the latter as they believe it is the greatest potential for inducing and sustaining flow. However, the main focus of Shute and her team is on stealth assessment, embedded within a serious game. Both epistemic and stealth assessments will be explained further under game models (see section 3.2.3).

thoughts, attitude, interaction patterns, nonverbal behaviour, facial displays, and silence (Bente & Breuer, 2009, p. 331).

Wouters & van Oostendorp (2017) proposed a list of guidelines for the use of structural assessment[18] within serious games. Their guidelines to use structural assessment in serious games include: determining the appropriateness of the domain for such assessment; selecting an appropriate referent for the target group(s); selecting the number of concepts regarded important in a domain; and analyzing the graphical knowledge representations to obtain in-depth information about the quality of the knowledge structures. Following this study, Wouters & van Oostendorp (2017) offered nine instructional techniques to facilitate learning and motivation of serious games[19] , in their meta-analysis.

GBA vs. Gamification. As mentioned above, GBA uses serious game structures for educational purposes, which makes it different from gamification. Gamification, defined as "the use of video game elements in non-gaming systems to improve user experience and user engagement" (Deterding, Sicart, Nacke, O'Hara, & Dixon, 2011, p. 1), has also become a popular technique across a variety of contexts. Gamification involves the use of game mechanics (points systems, badges, and so on) outside of a game, in a pre-existing process (e.g., classroom or training programme) to increase participation and engagement. On the other hand, GBA incorporates a mixture of game elements within the context of a game and involves educational values and positive reinforcement beyond entertainment and motivation (Marczewski, 2013). Therefore, GBA in this paper refers to the use of serious games with the purpose of developing and measuring certain skills, aligned to learning outcomes.

3.2.1 *GBA and 21st century skills*

Many skills like collaboration, decision-making under pressure, calculated risk-taking, critical thinking and ethical behaviour can be measured by GBA. O'Neil, Wainess, and Baker (2005) identified five types of cognitive demands in video games: content understanding, problem-solving, self-regulation, communication and collaboration/

18. A structural assessment to measure the quality of knowledge structures comprises three steps: knowledge elicitation, knowledge representation, and knowledge evaluation. In this way, structural assessment may add a deeper understanding of important concepts in a particular domain (Wouters et al., 2011). This is done by having individuals rate the relatedness of pairs of concepts.

19. Content integration, context integration, assessment and adaptivity, level of realism, narration-based techniques, feedback, self-explanation and reflection, collaboration and competition, and modelling.

teamwork. Constructs such as persistence or creativity can be assessed through digital game interaction (DiCerbo, 2017). The same is true of systems thinking (Mislevy et al., 2014), scientific inquiry and critical thinking (Gobert, Kim, Sao Pedro, Kenned, & Betts, 2015), and problem-solving in a number of different areas and contexts (e.g., Chang, Wu, Weng, & Sung, 2012; Graesser, Dowell, & Clewley, 2017). Furthermore, game activities are essentially performance-based as they involve problem-solving and challenges that provide players/learners with a sense of achievement (Prensky, 2001). It would facilitate both the development and assessment of hard-to-measure 21st CC in a more prolific and meaningful way compared to a traditional test (Gee, 2004; Squire, 2011; Shute, 2009). Similar studies on *Plants vs. Zombies 2* (Shute et al., 2017) and on *Oblivion* (Shute et al., 2009) claimed that these games can be a natural medium to assess problem-solving and creativity skills. Torres (2009) recently reported on his research on the game *Gamestar Mechanic*, tailored to teach kids basic game design skills so that they can actually build their own games. Based on the research findings, the children who played the game developed systems-thinking and innovative design skills.

However, Connolly et al. (2012), through an extensive literature study on GBA, asserted that despite a wide range of positive learning and performance impacts in GBA, there is limited evidence and lack of adequate measurement tools to methodically examine 21st century skills, and more research is required in this area to ensure that what is expected to be measured is really measured (Bente & Breuer, 2009). Thus, the cognitive and non-cognitive or affective dispositions of a competence must be clearly distinguished and separately measured. Zapata-Rivera & Bauer (2012) also warned against unexpected knowledge, skill or behaviours that are not part of the targeted skill but might occur during the game process. This might lead to the issue of construct-irrelevant variance. There are different sources of construct-irrelevant variance in the process of GBA. Unexpected interactions in a game environment, game features (e.g., actions across time points, cognitive loads or demands), students' verbal or IT proficiencies are some of the most common sources. For example, students' lacking background knowledge, not knowing how to use the interface, and not knowing what is expected can cause some difficulties to assess the skills which are expected to be measured. Even students who are good

in a content area may exhibit poor performance due to lack of familiarity with game play.

On a related note, game features increase autonomy and engagement but from an assessment perspective, increased autonomy can decrease the comparability of evidence across players (Mislevy et al., 2014). The influence of screen size, screen resolution and display rate on performance in both CBA and GBA cannot be overlooked either.

In a study, Sourmelis, Ioannou, & Zaphiris (2017) presented a literature review of MMORPGs (Multiplayer Online Role Playing Games) empirical research from 2010 to 2016. Although Their findings suggested that MMORPGs are suitable environments to foster a variety of 21st CC, most of the MMORPGs research in their study focused on probing communication and collaboration skills while other skills such as creativity, problem-solving and information literacy were unexplored.

3.2.2 *GBA & assessment purposes*

The literature in the area of GBA suggests that digital games can be applied to the design of innovative assessment tasks for both formative and summative purposes (Zapata-Rivera & Bauer, 2012). One example for summative purpose is the game Startup, challenge 1–6, when the final score is given after completing each quest (in Hainey et al., 2015). Chang (2010) designed a game to teach and assess Java programming for both summative and formative purposes. The game was built on web browsers with AJAX technology, enabling students to play it across different platforms, and providing them with rapid response and instant interactions. Allen, Seeney, Boyle, & Hancock (2009) used the game *Infiniteams Island game (TPLD)*, for both formative and post-test summative in the form of questionnaires. The goal of the game was to assess players' team working abilities. Through the questionnaires of 240 students, they found that the players gained self-awareness about their skills through the game. Overall, the most common methods to assess students' final knowledge at the end of a game consist of post-tests surveys, tests, questionnaires or teacher evaluation.

Although there are a few examples of GBA for summative assessment as mentioned above, there is a stronger claim in the literature that the natural strength of GBA is in assessing the information that learners

naturally leave behind in the process of game playing. Such data can provide an indication of the learning progress for both the learner and teacher, pointing to formative assessment. Bente and Breuer (2009) claimed that serious games are ideal for formative assessment and in-game feedback because they have the opportunity to take advantage of the medium itself to provide continuous feedback in the form of less intrusive and more seamless formative assessment. Shute et al. (2009, 2013) and Eseryl, Ifenthaler and Ge (2013) pointed out that in an effective GBA, the assessment process is "embedded" as a part of the serious game and optimally integrated into the tasks, thus not interrupting the game flow. Therefore, the term "embedded" or "stealth" assessment (Shute, 2009) refers to the formative purpose of obtaining accurate information about the player inserted into the game or curriculum in an unobtrusive way.

Bauer et al. (2017) provided some important parallels between the formative assessment principles and design elements in games and asserted that game-design elements are a natural fit in the service of formative assessment principles. To prove their claim, they used an example of *Mars Generation One™ (MGO)* (a game developed by GlassLab in collaboration with NASA and the National Writing Project) to assess students' argumentation skills. There are three dimensions of argumentation competency addressed in this game: *Identify* components of an argument (e.g., selecting a claim and identifying evidence to support a claim); *Organise* the components to construct a coherent argument that supports a position; and *Use* the argument to debate with another position. The game is set in the near future on Mars. Players take the role of visitors from Earth and the first person not born on Mars to join the Mars school, Argubot (robot like creatures) Academy. In the narrative, there are 33 issues (missions) in which students can take positions and argue their cases. Teachers are also provided with additional instructional resources and feedback throughout the game. Feedback is on a daily basis in the form of an automatically generated "watch out" list of students who skip important aspects of the play (e.g., not collecting any data for claims during the explore phase which would prevent them from making any progress) and a "shout out" list of students who have made significant achievements in the game. It also provides immediate in-game feedback to students. For example, when a student pairs data

to a claim, the game reports back on the strength and relevance of the claim-data pair, or the option for creating a stronger claim. Students cannot advance to a harder mission until they have successfully completed the prior level (Bauer et al., 2017).

Another instance of integrating formative assessment into a game is from Weng, Fakinlede, Lin, Shih, and Chang (2011), in the personalised game *QuizMASter*. The QuizMASter supports formative assessment by providing adaptive testing and feedback using intelligent pedagogical agents who pose questions from the repository and then provide adequate feedback to the players and instructors. Players are assessed on their ability to answer questions correctly. The game enables instructors to use assessment as a motivational tool within an entertaining context so students will feel less test anxiety.

Delacruz's research (2011) also supported the formative benefits of GBA while examining the impact of varying level of detail about a game's scoring rules on learning and performance in mathematics. She found that combining elaborated scoring explanation with incentives for accessing game feedback resulted in higher learning performance. Another example is from Graesser et al. (2010). They explained that *Operation ARIES*, a tutorial system with a formative assessment component for high school and higher education students, is designed in a game environment to teach and assess critical thinking in science. The system includes an *Auto Tutor*, whereby animated characters interact with students, interpret their response, and respond in a way that is adaptive to student reply. A similar research study by Graesser, Dowell, & Clewly (2017), aimed to assess collaborative problem-solving, utilising *Auto Tutor* and *Trialogues*, which are the conversations between players and two computer agents. *SimScientists* (Quellmalz & Pellegrino, 2009) was also developed for formative assessment for students around the age of 12 and comprises a virtual environment in which students can engage in science tasks.

In short, many game-based assessments are being used or developed more for formative purposes. Perhaps, the inherent abilities of the games and the increasing significance of formative assessment in the process of teaching and learning make GBA more suitable for

the in-process assessment than the usually high-stakes summative assessment. With the psychometric challenges improving in GBA (e.g., ECD framework developed by Mislevy et al., 2014), coupled with rigorously validated and highly engaging games, we might see it being more integrated for both summative and formative purposes in the future (Kato & de Klerk, 2017).

3.2.3 *Game models for assessment*

Two game models or designs to employ for an effective GBA are discussed here: stealth assessment and epistemic game models.

Stealth Assessments. When assessments are seamlessly embedded into the fabric of the game environment and are virtually invisible, it is called *stealth assessment* (Shute, 2009; Shute & Ventura, 2013). Such assessments, mostly for formative purposes, are intended to support learning, maintain flow, and remove (or reduce) test anxiety without sacrificing validity and reliability (Shute et al., 2008). Assessment tasks in stealth assessment are supposed to be so interactive, engaging, and immersive that players lose sense of time and are not conscious of being assessed. Data collected in stealth assessment can be used to make meaningful interpretations about learner performances and present a wide range of evidence and arguments to measure their achievements (Bellotti et al., 2013). Thus, stealth assessment within serious games offers the opportunity to inform and support a wider variety of knowledge, skills, and thinking needed for the 21st century education (Shute et al., 2009), and employs a principled assessment design framework called *evidence-centred design* (ECD) which integrates assessments directly into learning environments (Shute & Ventura, 2013). Shute referred to the stealth assessment as a "specialised implementation of ECD" (Shute et al., 2017, p. 62).

ECD framework (Rupp et al., 2010; Mislevy et al., 2012; Almond et al., 2003) uses the sequence of actions and behaviours as evidence to assess students' knowledge and competencies. The objective is to provide adaptive gaming scenarios (including various interactions or student behaviours and decisions) to navigate the task that can be used to provide valuable feedback to students and teachers. The connection between learning, behaviour, and setting provides support for the validity of what is being assessed (Shaffer & Gee, 2007). That

is why ECD is a bridge between the two demands for validity and rich assessment data that GBA have to address (Rupp et al., 2010). An important aspect of the ECD framework, such as two examples from the Cisco Networking Academy, *Packet Tracer* (as a simulation-based assessment) and *Aspire* (as a game-based assessment), is the seamless integration of learning tasks and assessment, which allows the collection of evidence based on students' performances and behaviours throughout the game. Walker and Engelhard (2014) argued that using ECD in GBA provides an opportunity to focus on students as multidimensional individuals acting within a particular context, making it relevant and suitable for assessing 21st century skills.

ECD consists of competency, evidence, and task models which provide "a framework for developing assessment tasks that elicit evidence (scores) bearing directly on the claims that one wants to make about what a student knows and can do" (Shute et al., 2009, p. 6). The competency model defines knowledge, skills, and other attributes that need to be measured in the game. The evidence model would show, how and to what degree student performance on particular tasks or game activities can be used as evidence to make inferences about their levels of competency. The task model specifies the activities or conditions under which data are collected. Task model designs must include the characteristics that will affect validity and difficulty of the activities, operational limitations, and game-related aspects that will keep students engaged in the game (e.g., immediate feedback and progress indicators) (Mislevy et al., 2012).

Different research studies revealed how stealth assessment coupled with ECD can provide a viable solution to assess skills, such as problem-solving or reasoning skills demonstrated during a game session (Shute et al. 2017; Shute & Ventura, 2013). Shute et al. (2017) and Shute, Masduki, & Donmez (2010) demonstrated in nine steps how to design and embed a stealth assessment combined with ECD within a commercial game in order to examine relevant knowledge and skills educationally:

1. Develop competency model (CM) of the targeted skill (based on the literature review)
2. Determine the game to embed the assessment in

3. Delineate a list of relevant game actions/indicators as evidence to inform CM
4. Create new tasks in the game (task model)
5. Create Q-matrix (Almond et al., 2003) to link actions/indicators to targeted skill
6. Determine scoring rules as parts of the evidence model, using discrete categories (e.g., yes/no, very good/good/ok/poor relative to quality of the actions).
7. Connect each indicator to the related CM variable(s), and establish a statistical relationship between them
8. Pilot test Bayesian Networks[20] to modify the difficulty and discrimination parameters accordingly
9. Validate the stealth assessment with external measures.

Following the steps above, Shute el at. (2017) developed and embedded a problem-solving stealth assessment into an existing game called *Use Your Brainz* (a modified version of the game *Plants vs. Zombies* 2). In the game, players position a variety of plants in front of their houses to prevent zombies reaching them. Each plant has different attributes: offensive plants attack zombies directly, while defensive ones slow down zombies to give the player more time to attack. The challenge of the game comes from determining which plants to use and where to position them on the battlefield to defeat all the zombies in each level of the game. The authors explained step-by-step creating and implementing the stealth assessment of the problem-solving skill through the competency, evidence, and task models they created in this game. They identified relevant indicators in the game to provide evidence of players' levels on the competency model, made scoring rules for each indicator, and connected the indicators statistically with competency model variables. They then modified the Bayesian networks by collecting and analysing data collected from a pilot study. They examined the validity of the test by selecting some well-established external measures such as Raven's Progressive Matrices (Raven & Court, 1998) and MicroDYN[21] (Greiff & Funke, 2009) to measures problem-solving skills.

Another example is from Shute & Ventura (2013), which presented

20. A Bayesian network is a probabilistic graphical or statistical model to represent a set of variables and compute their relationships (en.m.wikipedia.org).

21. Proposed by Greiff & Funke (2009), MicroDYN is a computer-based approach to assess collaborative problem-solving. This approach was developed under a psychometric perspective for the use in large-scale assessments to search for a complex system.

the stealth assessment of creativity skill within a game called *Physics (Newton) Playground*. *Physics Playground* is a two-dimensional physics game designed to assess players' non-verbal understanding of physics, persistence, and creativity. The goal of the game is to create objects on a screen using a mouse and coloured markers to help the green ball reach the red balloon through utilising the laws of physics such as Newton's laws of motion, mass, gravity, conservation of energy and momentum. To illustrate how the creativity stealth assessment works in this game, the authors explained one of the game's 74 levels called "Swamp People" (a medium-difficult level) that they tested with 167 middle school students. They discovered that the most common solution used in that level was to create a ramp from ball to balloon. Less frequently, students created a springboard to solve the level. In just one case, a student used a lever to solve the problem which provided positive evidence for flexibility and originality, as two aspects of creativity. In this way, different behaviours and decisions during the game can be used to infer students' level of creativity. Each level in the game had its own Bayesian network as the levels differed in terms of difficulty. When students successfully solve the problem or leave the level, data from the log file would be unobtrusively gathered and analysed; indicators would automatically be created, scored, and inserted into the Bayes net to infer students' level of creativity either in one level or overall in the game. Kim and Shute (2015) confirmed in their study with 167 middle school participants (grades 8-9) that *Physics Playground* is a valid and reliable tool for GBA.

Epistemic games. *Epistemic games*, developed by Shaffer and his team at the University of Wisconsin-Madison (Shaffer & Gee, 2007), are another innovative model for GBA for 21st century skills, allowing learners to develop domain-specific expertise, based on principles of collaborative learning and complex problem-solving, for real-life professions (Bagley & Shaffer, 2009). Epistemic games are computer games where players learn to think and act like professionals—such as journalists, artists, managers, or engineers. Epistemic Network Analysis (ENA) translates the elements of ECD in the game into a knowledge network map. As such, ENA tracks players' competency levels through the game, which can be continuously quantified, analysed, and updated to assess their progress and to inform selection of game task and

activities to be presented (Shaffer, Collier, & Ruis, 2016).

Epistemic games make both individual and collaborative experiences accessible to learners, characterised in real-life and virtual settings. Notably, the objective of epistemic games is not to train learners toward specific career trajectories, but to facilitate and transfer disciplinary thinking as well as core competencies (Rupp et al., 2010). Learners are expected to think critically, make decisions in the game under real-life constraints and in real-time, and solve related problems (Bagley & Shaffer, 2009). For example, *Legends of Alkhimia*, developed by Chee and his team in the National Institute of Education, Singapore, was designed to foster the learning of chemistry through inquiry for 13 to 14-year-olds. In this multiplayer game, students take the role of chemists to solve challenges related to the use of chemistry in realistic contexts (Chee, Tan, Tan, & Jan, 2012). Epistemic games are currently pushing the methodological boundaries of educational assessment, and are still at the infant stage. More development in studies and narratives is required to warrant what is claimed[22].

Using the epistemic game *Urban Science* (Bagley & Shaffer, 2009), the authors illustrated the numerous decisions that need to be made during game development and implications for accumulating qualitative and quantitative evidence about learners' progress during the game. *Urban Science* is based on the professional experiences of urban planners. The game (see http://epistemicgames.org/eg) is continually updated at the University of Wisconsin and used as the leveraging environment for the NSF-funded *AutoMentor* and *Dynamic STEM Assessment* grants, whose goals are to develop automated feedback mechanisms for epistemic games. In this game, students need to use technology, scientific understanding and communications to develop innovative solutions for real problems. They use a special mapping tool, called iplan (which is a tool similar to an actual Geographic Information System), to come up with their final planning. Towards the end of the game, they write their final proposal to the mayor discussing the strengths and weaknesses of their solutions. While playing the game, the relevant skills, knowledge, values and epistemology from urban planning can be assessed by observing what players say and do in the game, creating a model of their problem-solving process during

22. A more comprehensive overview of epistemic games can be found in Shaffer & Gee (2007).

the game, and finally comparing that to how professional planners think (Shaffer et al., 2016). In the current version of *Urban Science*, all tasks and most communication are accessed through a web portal of the fictitious company *Urban Design Associates*. The learners receive e-mails from virtual agents that are represented by photos of real people. The e-mails describe the tasks that they have to perform and the resources available via the portal. The tasks are a mixture of real-world tasks (e.g., learners travel to the actual neighbourhoods in Madison to take photos and notes) and virtual-world tasks (e.g., learners research information about background reports and write their own). In *Urban Science*, much of the relevant learning experience happens through collaboration and cooperation as students are expected to be able to communicate and negotiate clearly—orally and in writing—to collect, organise, and analyse information; therefore, it can be a suitable platform to assess communication and collaboration skills.

Epistemic games include highly complex tasks, based on which reliable assessment data (skills, knowledge, identity, value) are collected to develop the epistemic frame (Rupp et al., 2010). Therefore, it is highly educational but costly and resource-intensive. Since many tasks in epistemic games are collaborative, the role and degree of scaffolding and mentoring (by teachers) is critical as it influences student interaction (with themselves and mentors) as well as their contributions toward task completion. The manner by which mentors interact and engage students can influence students' thoughts, actions, and rationales and consequently the quality of the observable data to be collected (Rupp et al., 2010). However, quantifying reliability and minimising measurement error in the assessment design are very challenging due to the complexity and richness of the tasks. To address the reliability and validity measurement, the ECD framework (student, task, and evidence models) is suggested as it identifies different layers where various activities take place (Rupp et al., 2010).

In summary, section 3 has sought to provide some insight into technology-based assessments, particularly GBA, and the way they can facilitate learning knowledge and promoting 21st century skills both inside and outside of schools if they are designed and developed properly. Indisputably, achieving meaningful assessment through

designing and exploiting digital games will be challenging for the 21st century educators and serious game developers. Some of these challenges will be pointed out in the following section.

Section 4: Challenges, Implications, and Future Recommendations

The move towards holistic learning and developing 21st century skills requires a comprehensive system of assessment. However, assessing 21st century skills is multi-faceted and difficult to condense, given the multidimensionality and expansive definitions of the skills. There is a great deal of research and policy work concerned with changing the standardised testing system and building more complex and authentic forms of assessment; however, all suggest that any change can be contentious and challenging (Lederman & Abell, 2014). The advantages and limitations of the potential methods to assess 21st CC have been pointed out in section 2 of this paper. This section summarises some general challenges plus the relevant implications and then presents some suggestions for future policy-making and research.

4.1. *Challenges & Limitations*

Conceptual challenges: The first challenge in assessing 21st century skills is conceptual which lies in the lack of a comprehensive and definitive understanding of the nature and development of some of these skills (especially in relation to non-cognitive or affective skills), and the difficulty to partition variance in behaviours attributable to knowledge or attributable to skill. Conceptual challenges are largely related to grappling with new skills or skills with a broad spectrum of definitions, causing overlapping definitions of each facet or construct. For example, collaboration, coordination and cooperation might be used synonymously but are actually different concepts (Koenig, 2011). Relatedly, the complexity of some skills poses additional challenges to measure them. Skills such as collaborative problem-solving or global citizenship are referred to as complex skillsets (Care et al., 2018; Funke et al., 2018) as they draw on a combination of cognitive and non-cognitive constructs as well as values, knowledge and attitudes. It is sometimes difficult to separate these aspects and to find a careful selection of tools to adequately measure them (Care & Kim, 2018). To address this challenge, Funke et al. (2018) proposed the use of technology-based assessments that can promise a more

comprehensive measurement, facilitating a plurality of assessment instruments to assess different aspects of a targeted skill. However, the challenges and barriers associated with TBA cannot be underestimated.

Technology-related challenges: Despite all the potential benefits and innovative opportunities offered by TBA, and more specifically by GBA (as elaborated in section 3), there are barriers to adopting this form of assessment in formal education settings, mainly related to getting the assessment right. The technological and practical barriers cannot be overlooked (Griffin & Care, 2015). Lack of technical expertise or limitation of knowledge to identify and troubleshoot problems during the assessment would be an added challenge. Time and cost are the primary critical concerns (O'Neil et al., 2003). To determine whether a serious game is successful in achieving the intended learning outcome and assessing the appropriate knowledge, skills or abilities is a complex, time-consuming and expensive process (Bellotti et al., 2013).

Another important concern is creating engaging scenarios for assessment purposes while satisfying assessment requirements—fairness, validity and reliability (Zapata-Rivera & Bauer, 2012). Opponents of GBL and GBA argue that games are just another technological trend, leading to superficial learning only. They argue that games may cause violence, aggression, inactivity and obesity while decreasing prosocial behaviours (Gentile et al., 2009). Technological advancement highlights the changing nature of communication and collaboration skills. Generations of *digital natives* (Prensky, 2001) have different views of social norms and different ways of interaction and communication that occur less face-to-face and more face-to-screen. The challenge is to interpret these evolving patterns and the quality of students' thinking as they respond to them.

Another major limitation of using ICT for assessment functions, especially in large-scale assessments, is the lack of uniformity. Providing a uniform testing environment is a problem when school IT facilities and infrastructures can vary considerably. The influence of variations in screen size, screen resolution, and the internet speed and traffic on performance is also significant (Bridgeman, Lennon, & Jackenthal, 2003). In practice, a high capacity of internet connection is required for a specified number of students completing the assessment

simultaneously. Consequently, it is necessary to cautiously consider the connection speed to guarantee successful online assessment delivery.

Besides validity and reliability, security issues may also affect the assessment process in TBA, depending on the purposes and contexts of assessments. Security, confidentiality, and integrity of tests, personal data and results must be ensured at a server level. In the case of internet-based testing, various risks such as hijacking of websites or virus attacks are critical threats (Csapó, Ainley, Bennett, Latour, & Law, 2012). Moreover, the issue of plagiarism must be identified and addressed.

Social-cultural factors: social, cultural, and family backgrounds can affect student performance. Some students may come to school better prepared to collaborate with others or are more comfortable in a certain assessment environment. This may occur as a result of family or cultural backgrounds. For instance, English and IT proficiencies or IT familiarity, which are highly cultivated by home environments, can affect some assessment tasks in the case that reading, verbal communication or IT literacy are required. Thus, the question is: To what extent would the selected assessment be measuring skills learned in school versus out of school?

On a related note, school leaders' and teachers' perspectives and attitudes towards innovative forms of assessment (e.g., GBA) in formal educational settings are an important challenge to consider. Any form of technology-based pedagogy requires teachers to be willing and prepared to engage with classroom pedagogy differently, and being different from what is normative inherently invites resistance and unfavourable response. Reports from ATC21S pilot studies confirmed that teachers' unfamiliarity and uncertainty regarding the 21st century skills and assessment techniques made them feel ill-equipped to provide feedback. One example is a report by Poon, Tan, Cheah, Lim, and Ng (2015), a part of ATC21S project conducted in Singapore assessing collaborative problem-solving through digital networks. They found that the new concepts of teaching and assessment paradigm such as open-ended or ill-structured assessment tasks and tracking and interpreting student behaviours in collaborative settings were "troubling" for teachers and unfamiliar for students (Poon et al., 2015, p.

199).

Chee, Mehrotra, and Ong (2014) illuminated teachers' dilemmas in their attempt to enact the *Statecraft X* game in Singapore Social Studies curriculum as one of the impediments. They identified four distinct situations which gave rise to teachers' dilemmas: discomfort with a new mode of pedagogy; teachers' perception of system requirements and normative expectations; pressure to ensure students' high marks on standard tests; and weak alignment between mandated and innovation-based forms of assessment. Findings of this and similar studies indicate that changing teachers' sceptical mind-set towards applying TBA/ GBA in classroom requires a more flexible curriculum and educational system in which institutional authority, practices, and values not being perceived to be perpetually challenged. When there is a proper alignment between curriculum, educational expectations, and game-based learning and assessment, games will be regarded more than just play but as a potential educational tool that could fundamentally change teaching and learning for the upcoming generations.

Overall, the literature on 21st century skills assessment suggests that there is still much work to be done to develop effective assessment processes with satisfying psychometric properties.

4.2 *Recommendations*

4.2.1 Recommendations for policy-making

To assess 21st century skills, it is essential to adopt an assessment process which is coupled with a clear definition of the targeted skill (including the granularity, domain specificity, etc.), careful specification of learning outcomes, and a detailed and well-communicated implementation plan. Such assessment must be authentic, rigorous and comprehensive so that it can cultivate students' active learning by posing complex problems and helping them solve these problems. Both summative and formative assessments with the emphasis on students' self-assessment are required. A move towards promoting assessment *of, for,* and *as* learning as parts of one process will provide opportunities for strengthening the teaching, learning, and assessing of 21st century skills (Darling-Hammond, 2012). In short, focusing on

a student-centred approach is the pre-requisite for any successful assessment system to prepare students for the world that awaits them beyond the school doors.

High quality teacher training programmes, including continual professional development processes, in line with the recommendations from the national and international literature, are essential to prepare teachers for 21st century skills assessments (Timperley, 2008). Teachers need to be trained and provided with incentives, resources and sufficient time for selecting appropriate and effective assessment tasks. Dedicating more time across all subjects for performance tasks with the sufficient level of difficulty to be carried out in the curriculum is vital. Similarly, the given tasks must be open-ended and ill-structured to allow students to use more metacognition and decision-making skills to solve the problem (Ku, 2009; Turner, 1995). To achieve this, it is helpful for teachers to have access to a repository of tasks, along with level-appropriate assessment rubrics, moderation processes and resources catered to the local education context. Rubrics based on the developmental progression of each skill across the primary and secondary levels should be provided and enriched. In short, if the goal of assessing 21st century skills is for students to acquire these skills and for teachers to be accountable for teaching them, precise instructional programmes are needed and details of fairness and validity must be investigated and understood before moving towards the wide use of assessments.

In consonance with 21st century skills teaching and assessment, instilling supportive attitudes in teachers is crucial because teachers ultimately exercise and implement them through their day-to-day practices including assessment (Bowe, Ball, & Gold, 2017). To change teachers' and school leaders' mind-sets towards change, personal and professional transformation is required which involves a difficult (but not impossible) process of shifting from a stable, habituated practice to a rearranged mode of 'enactive being' (Chee et al., 2014). Teacher readiness depends on the school's readiness to support the use of digital technologies in teaching and learning (Petko, Prasse, & Cantieni, 2018). Therefore, school leaders also have a significant role in fostering a positive mind-set towards new approaches in daily teaching and learning. To achieve this, educational administrators

and policymakers must understand and acknowledge the challenges that educators must navigate to transform successfully. Therefore, professional development training on how to design the assessment tasks, to understand the skills and their developmental learning, to administer the assessment, and to interpret the assessment data, is indispensable (Care et al., 2018). Wiliam (2007) identified the need for a gradual assessment approach, with each teacher implementing not more than two or three assessment strategies at any one time to avoid a loss of routine and confusion. Furthermore, he argued that the approach should be flexible because some techniques in one context may not work in others or need adjustment (see also Resnick, Spillane, Goldman, & Rangel, 2010).

Assessment systems should provide multiple methods that support the triangulation of inferences (as mentioned in section 1.1), which validate the assessment data obtained from different sources through cross verification. A number of mentioned challenges in measuring 21st century skills suggest the need for multiple measures to address the limitations of any single method. Different methods can tap different aspects of the same construct (Lai & Viering, 2012; Treffinger et al., 2002). Although researchers argue that education policies and practices need to exploit the potential of existing technologies through research, development and evaluation (e.g., Redecker, 2013), lack of familiarity in this field may negatively affect both teacher and student performance and motivation. Using supporting tutorials and demos can facilitate the familiarisation process and consequently increase engagement (Armstrong & Georgas, 2006).

With regard to technology and game-based assessment, it is worth-repeating that games and advanced technologies work best when coupled with effective curriculum and pedagogy (Squire, 2011). It is a valid concern that virtual environments could replace real-life interactions and that schools might lose sight of their ultimate objective, which is preparing young people to be well-informed and productive in the 21st century society. As such, games should not replace teachers and classrooms. In fact, direct interaction with teachers and peers is an important skill for young learners and teachers should ensure that technology is used only as an instrument to promote this skill.

4.2.2 Recommendations for research

Regarding GBA, there is an acknowledged need for high quality, cost-effective studies to support the use of serious games in educational settings (Girard, Ecalle, & Magnan, 2013). Research on GBA needs to analyse the flow of both the curriculum and the game to ensure that assessment tasks are tailored to different groups, and learners are equally and emotionally engaged in the process. When implementing digital assessments of certain skills such as collaborative problem-solving or ICT literacy, sufficient and equal accessibility to technology is required for such assessments to be even possible. In terms of group size, there should be definitive research concerning ideal group sizes and interactions. Most research recommended a group size of four or five in an educational setting (Johnson, Johnson, & Holubec, 1994), suggesting that the smaller the group size, the higher the chance for each member to participate actively.

Game designers should conduct a series of trials to understand how different game designs ultimately affect players' goals and behaviours within the game. Assessment designers should use the information obtained from these trials to iteratively build assessment mechanics (e.g., scoring rules, evidence accumulation methods) (Mislevy et al., 2014). One implication is that if the primary goal of GBA is learning, rather than assessment, it is advisable to provide more control to the players, allowing them to explore different strategies and solutions rather than achieving success quickly. GBA undeniably is an interesting field of research which is full of promise and challenge but requires deeper attention from both research and practice.

In the local literature, the relationships between long-term impacts of innovative and authentic assessments of 21st century skills and learning behaviours are lacking. Thus, future research should look at developing and testing multi-dimensional measures of 21st century beyond the use of self-report instruments and examine the relationship between 21st century skills and academic outcomes in Singapore schools. This might also help level up low-progress students academically.

Perhaps what Voogt & Roblin (2012) highlighted can be the summary of main recommendations for assessing 21st century skills: “defining

goals and standards in national documents, regulating the curriculum, embracing a powerful vision, encouraging collaboration between different sectors, building on already existing work and focusing on what is do-able, ensuring equitable access to education in present and future society, stimulating teacher collaboration, creating learning environments that enhance competence development, and aligning assessment methods and goals" (p. 312). Assessing 21st century skills is effective when policymakers, educators, game designers and assessment experts work together from the onset. This type of heterogeneous team is critical for the *transformative assessment* to fulfil the educational and economic expectations of our time.

Appendix 1

The following table shows the comparison between the most popular and frequently cited 21st century frameworks and their classifications of skills:

ATC21S (Binkley et al. 2012)	**UNESCO (2012)**	**OECD[23] (2005)**	**Partnership 21 (2009) www.p21.org.**	**European Commission (Gordon et al. 2009)**	**MOE of Singapore (2010)**
Ways of thinking: Creativity & innovation, Critical thinking, Problem- solving, Decision making, Learning to learn, metacognition	**Learning to know** **Learning to be**		**Learning & innovation:** Creativity, Critical thinking, Problem-solving Communication & collaboration	**Learning to learn:**	**Critical & inventive thinking:** Sound reasoning & decision making, Reflective thinking, Curiosity & creativity, Managing complexities & ambiguities Subjective; Halo effects; Rating scales and language load to be modified and simplified for young learners
Ways of working: Communication, Collaboration	**Learning to do**	**Interact in heterogeneous groups:** Relate well to others, Co-operate-work in teams, Manage and resolve conflicts		Communication in mother tongue & foreign languages,	
Tools of Working: Information literacy, ICT literacy		Formative and summative	**Information, media & technology**: Information literacy, Media literacy, ICT literacy	Mathematical, science & technology competences, Digital competence	**Information & Communication skills:** Openness, Management of information, Responsible use of information, Communication effectively
Living in the World: Citizenship-local and global, Life & career, Personal & Social, Responsibility-cultural awareness & competence	**Learning to live together**	**Act autonomously**: Act within the big picture, Form and conduct life plans & personal projects, Defend & assert rights, interests, limits & needs	**Life & career:** Flexibility & adaptability, Initiative & self-direction, Social & cross-cultural skills, Productivity & accountability, Leadership & responsibility	Social & civic competences,	**Civic literacy, global awareness & cross-cultural skills:** Active community life, National & cultural identity, Global awareness, Social-cultural sensitivity & awareness

23. The OECD's Future of Education and Skills 2030 project revisited the 2005 key competencies and found them to be still relevant today and in 2030. However, it articulated three inter-related and transformative competencies, required for the revised framework: Creating new value, Taking responsibility, and Reconciling tensions, dilemmas, trade-offs, and contradictions. The relevant constructs are still being reviewed, and revising the pedagogy and assessment process will be starting in 2019 (http://www.oecd.org/education/2030).

Appendix 2

Below is the summary table of some examples of 21st century skills being assessed by different methods:

Skill	Assessment Method	Assessment Name	Reference
Creativity	Self-report	Creative Domain Questionnaire (CDQ); Creative Behaviour Inventory (CBI)	Silvia, Wigert, Reiter-Palmon, & Kaufman (2012)
		Creative Achievement Questionnaire (CAQ)	Carson, Peterson, & Higgins (2005)
	Project based performance rubrics (for Grades K-2, 3-5, 6-8, 9-12)	Buck Institute for Education (BIE)	http://www.bie.org/ http://bie.org/blog/how_to_use_the_4cs_rubrics
	Performance task	Deeper Learning Educator Micro-credentials, Creativity	Digital promise annual report (2016): https://digitalpromise.org/wp-content/uploads/2016/03/2016-Digital-Promise-Annual-Report.pdf
	CBA	The Mission Skills Assessment	https://missionskillsassessment.org/ https://21k12blog.net/2014/09/05/mission-skills-assessment-toolkit/
	GBA	Newton's/Physics Playground	Shute & Ventura (2013)
Critical Thinking	Performance task	Collegiate Assessment of Academic Proficiency (CAAP)	ACT 2015: http://www.act.org/content/dam/act/unsecured/documents/ACT_Technical_Manual.pdf
		Halpern Critical Thinking Assessment (HCTA)	Halpern (1998)
		World Savvy Challenge	RAND & Asia Society, 2013, *Measuring 21st century competencies: Guidance for educators:* http://asiasociety.org/files/gcenmeasuring21cskills.pdf
	Project based performance rubrics (for Grades K-2, 3-5, 6-8, 9-12)	Buck Institute for Education	http://www.bie.org/
	GBA	EcoMUVE; SimScientists	RAND & Asia Society (2013)
	Multiple methods	Queensland Performance Assessment (QPA)	RAND & Asia Society (2013)
Communication & Collaboration	Questionnaire	The Competent Speaker Speech Evaluation Form (CSSEF) The Conversational Skills Rating Scale (CSRS)	National Communication Association (NCA), (2007, 2015): https://www.natcom.org/uploadedFiles/Teaching_and_Learning/Assessment Resources/PDF Competent Speaker Speech Evaluation Form 2ndEd.pdf
	Portfolio & Performance task	Singapore Group Project Portfolio	RAND & Asia Society (2013)
	Performance task	ACT Work Keys Teamwork Assessment Video-based	
	CBA & Performance	PISA 2013, 2015	OECD (2013)
	GBA	Alelo language and culture simulations	RAND & Asia Society (2013)
		SimScientists	
		EcoMUVE	
	Multiple methods	Queensland Performance Assessment	RAND & Asia Society (2013)
		PARCC/SBAC Common Core Tests Communication Items	http://www.parcconline.org/assessments/testdesign/research
		Mission Skills Assessment (MSA), Collaboration	https://missionskillsassessment.org/

Appendix 2

Skill	Assessment Method	Assessment Name	Reference
Problem- solving	CBA & Performance task	MicroDYN	Greiff & Funke (2009)
		ATC21S	Griffin & Care (2015)
		PISA 2012	OECD (2013)
	GBA	McLarin's Adventures, (MMOG)	Eseryel et al. (2013)
ICT literacy	Questionnaire	Different case studies	Chu et al. (2012)
	CBA	IEA International Computer & Literacy	Fraillon et al. (2015)
		ATC21S	Griffin & Care (2015)
Global awareness/ Citizenship	Questionnaire	The Interpersonal Reactivity Index (IRI)	Davis (1983)
		Southeast Asian primary learning metrics	Fraillon et al. (2015)
	Performance task	World Savvy Challenge	RAND & Asia Society (2013)
	GBA	REAL LIVES	Bachen, Hernández-Ramos, & Raphael (2012)
		Global Conflicts	http://www.globalconflicts.eu

Examples of Research Studies on GBA

As previously elaborated, today's students, more than any other time, are required to be equipped with skills such as technological fluency, complex communication and the ability to work effectively with others. Games can facilitate the creation of such environments where students can acquire and demonstrate these skills. Recent studies have started exploring how GBA can provide more details to assess these skills by highlighting its flexibility to easily keep track of players' moves and behaviours to measure targeted skills across a variety of domains (Shute et al., 2010). In what follows, some of these studies will be discussed, summarising their findings, implications and limitations.

** Quest Atlantis: Taiga Park*

Quest Atlantis is a multiplayer game with online and offline learning activities for children aged 9–15 where they have role play and make decisions in fictional circumstances. The activities are based on scientific enquiry, such as taking water samples and conducting interviews, and assessment is embedded in the activity. To solve the problem, the children need to develop knowledge, skills and attitudes aligned with scientific competence as well as with civic, social, and digital competencies (Redecker, 2013). The game can provide complex holistic problem-based environments, supporting active learning through authentic collaboration and immediate feedback (Squire, 2011).

In their study, Shute et al. (2010) described an approach to develop a stealth assessment of systems thinking and problem-solving competency, embedded in *Quest Atlantis*: *Taiga Park*, (developed by Barab, Sadler, Heiselt, Hickey, & Zuiker, 2007). To solve the problems in Taiga (a beautiful virtual park) and save the park, students should interview different characters in the park and hear their opinions about the cause of this problem, then write and submit short essays to their teachers as a required part of the missions. In addition to the essays, students can create and submit diagrams (demonstrating the relevant and existing variables and their cause-effect relationships) that can be uploaded as an attachment to their essays. Diagrams (created in jMap) are the platform for formative assessment and comparing students' diagrams and consequently the targeted competency. The use of ECD (Evidence Centred Design) as the embedded diagnostic system permits teachers, students or parents to examine the evidence and the competency levels and provide valuable feedback to the learner.

Shute and her team used the *Taiga Park* game, with its requirement for socio-scientific inquiry as well as continuous reflection and revision of current understanding, as an ideal environment to demonstrate the use of ECD and stealth assessment to assess student problem-solving skills. Other skills, such as teamwork and communication, can similarly be assessed in this game when a competency model (CM) has been developed and indicators fully identified (Shute et al., 2010). One of the challenges highlighted in this research (also addressed by Rupp et al., 2010) was developing the CM at an appropriate and precise level of granularity in the assessment. Too large a grain size means less specific evidence is available to measure student competency, while too detailed a grain size means more complexity and resources to deal with, for the assessment. Another challenge was in monitoring and scoring students' essays and online discussions. Teachers' overloaded schedules as well as some level of subjectivity were observed even when they were provided with comprehensive rubrics. Thus there needs to be a detailed and robust coding scheme to take into account the context of the tasks and semantic expressions in students' works. Alternatively, instead of spending countless hours grading essays and diagrams, teachers could simply review students' CMs provided by a software, and use that information as the basis for instructional and formative purposes. However, this is only appropriate for certain tasks (Shute et al., 2017).

In a similar study, Zuiker (2012) compared two Singapore secondary schools' use of the *Quest Atlantis: Taiga Park* in the class with each other, and with the previously published US research (Barab et al., 2007), to assess collaboration and communication. The participants were 39 and 36 third year secondary students in each class and both teachers were in their third year of service after completing a post-graduate certification program. Peer and class interactions as well as semi-structured student interviews and unstructured debriefing conversations with teachers were captured by audio/video records. A database documented student activities such as logins, chat entries, and written responses. Students completed both pre- and post-assessment, consisting of five multiple-choice items and two constructed-response items on problem-solving. Both classes used the as a resource for project or inquiry based activities. School A used it as a curricular and exam-driven frame, and school B as an extra-curricular and project-based one.

Results, collected from statistical inferences, descriptive accounts, and students' essays, showed general improvements and collective achievements in both schools. Students expressed enthusiasm about the experience.

Almost all continued using the virtual environment after the class, and 40% voluntarily logged in from home two or more times over a 3-week span. Classroom observations documented that interaction revolved mainly within-pair discussions on navigation (e.g., locating a character) and the evaluation of textual information (e.g., summarising a character interview). Significant content learning and teachers' positive evaluations of student products suggested meaningful individual engagement. However, the general patterns and negative cases indicated the disjointed and irregular peer interactions. The investigations during the game provoked minimal substantive collaboration skills and collaborative inquiry within or between pairs, and revealed little about the social aspects of learning. Thus, even though both classes engaged in information sharing, synthesis, critique and collaborative inquiry, the quantity, frequency and extent of student engagement remained unclear.

Compared to the previous US study (Barab et al., 2007), the Singapore study highlighted different cultural and situational perspectives. The biggest difference observed was the analytical stance which was increasingly common in the US study but consistently rare in both Singapore classrooms. The reason for this difference was the result of different types and definitions of interaction as the narrative designed by Barab et al. (2007) framed peer interaction in terms of efforts to enhance an analytical stance and it was more facilitated by teacher's efforts. This revealed that interaction operated more broadly than just the direct relationship between teacher and students. While using *Taiga Park*—in curriculum and GBA—challenges the conventional classroom dynamics and participation, it also illuminates the cultural boundaries and their influence on interaction, engagement, socio-scientific inquiry, teacher facilitation, and use of technology to transform conventions. Understanding and evolving the interaction among elements of a virtual environment and classroom conventions are undeniable challenges for educators and game/assessment designers, yet presenting a unique affordance of learning and assessing.

Therefore, games like *Taiga* could offer an opportunity to enact a collective investigation in a system where collaborative assessment is not mainstream, and consequently engage and evolve cultural boundaries. In short, the research suggested that circumstantial, social and cultural conditions underlying the learning and teaching processes and the mutual influences of each on the designed curriculum and assessment processes must be taken into account.

** Assessment of 21st Century ICT Skills In Chile*

In their project, Claro et al. (2012) evaluated 1,158 15-year-old Chilean students' ICT skills through a performance-based assessment within a virtual-game environment. This research was conducted in line with the Chilean Ministry of Education's policy, called Enlaces (Links), to improve the equity and quality of the education system by building a national network to support learning and develop ICT skills in schools[24].

The research team developed a software to simulate ICT applications and designed tasks to create real-life situations, including a virtual desk, email administrator, Internet browser, text processor, spreadsheet, programme for presentations plus a chat window where a virtual conversation among students was simulated. The assessment comprised three tasks, each related to one of the defined dimensions: Information (sourcing for and processing information to generate new information/ideas), Communication (interacting and contributing within a group and transmitting information effectively), and Ethics and Social Impact (responsible use of ICT legally, ethically and culturally at both personal and social levels). To create a flow of activities between the tasks, a game-like narrative on the theme of ecology was defined, presenting different situations in each task to students, by which their ICT skills could be tested. In the first task, students were assigned different activities to participate in a campaign for the protection of species near extinction. As for the second task, students were required to prepare a working document about global warming. In the third task, students were expected to identify risky behaviours in virtual environments and analyse the impact of the Internet on life. To do these tasks, students were required to discuss with their classmates and communicate their conclusions in specific settings like posting their ideas in a forum.

Majority of test items were designed in multiple choice questions but some would ask students to use specific tools of the simulated environment prior to selecting the answer (for example, reviewing emails to decide on the suitable tasks or searching for information about a particular topic on webpages). The test also included open-ended questions in which students were asked to execute certain functions such as posting an idea in a forum, reviewing or editing a graph or document, sending emails, and so on. Through these activities, the study aimed to examine how students learned to use ICT

24. For more information: http://www.unesco.org/education/lwf/doc/portfolio/abstract7.htm
Following this policy, by 2010, 90% of students in public schools had access to computers, 60% of schools had Internet access, 110,000 teachers were trained to use computers as a part of their instruction process, and the country reached a national average of 9.8 students per computer in schools (Claro et al., 2012).

and their perspectives on their own capabilities with regard to different ICT tasks. For instance, in the communication dimension, students were tasked to choose the best way to communicate a message while considering the purpose and recipient. They were to open and read five different draft emails, and then send the best one to a university professor, or select the best answer for a question in an ecologic forum, or pick the best slide to explain global warming to little children. To evaluate the ethics and social impact, students had to write a short essay to reflect on and discuss the implications of the Internet pertaining to different aspects of society.

Students' characterisation questionnaire was administered to collect data about their sociodemographic variables (individual and contextual), ICT access and use, and self-perception of ICT skills. The assessment explored how the variability of students' scores could be explained by their differences in terms of socio-economic group (SEG), gender, access and frequency of ICT use at school and at home, and confidence in doing ICT tasks. The findings showed that access, SEG, frequency of ICT use at home and confidence in doing basic ICT tasks were statistically significant predictors of students' scores. The factor confidence in basic ICT tasks was found as the highest percentage of variability in test scores in this study.

The results suggested that the test measured a combination of higher-order thinking skills, classical cognitive skills and ICT functional skills, and explained that the majority of students (three quarter) were able to solve tasks related to the use of information as consumers and search for information. Half of them could also organise and manage digital information. However, very few students (only one third) were able to complete tasks related to the use of information as producers to develop their own ideas in a digital environment. Less than one fifth of students could refine digital information and create a representation in a digital environment.

As for the limitation of this study, the authors believed that the assessment process was too long (2 hr 30 min) to cover three intensive skills. Although most tasks were performed collaboratively, collaboration skills were only evaluated at a declarative level, limiting the possibility to draw conclusions regarding students' actual skills. The test did not register students' ICT functional skills separately, limiting the analysis and the ability to report independently about these skills. As a part of students' characterisation questionnaire, information about their families' cultural capital (e.g., parents' level of education) was not registered.

In short, the test worked well and presented good psychometric properties; however, the study recommended schools to equally prepare students to apply ICT skills more as producers than consumers, and educational policies to develop actions to reduce digital divide (depending on students' SEG features) in the learning and assessing processes. It is essential to delve deeper into influential factors on ICT skills, such as the specific role of basic cognitive skills, the critical level of students' frequency of use and experience with digital culture, and the particular ICT uses and pedagogical practices that foster these skills.

** Zoo U (www.ZooUgame.com)*

Zoo U, a web-based strategy game was produced by 3C institute in North Carolina by a group of Ph.D. psychologists and game designers (DeRosier, Craig, & Sanchez, 2012). It aimed to engage students in a virtual school for zookeepers and assess their social as well as problem-solving skills. In this virtual school, children learn how to be zookeepers by completing some social problem-solving tasks to do with caring for the animals. The game combined theory-driven content and customized game mechanics and created the opportunity for stealth assessment, whereby players' choices and actions during gameplay provided the required data in real time as the evidence to assess social skills. The game has been researched and reviewed to have moderately good results for reliability and validity, correlated with school outcomes, GPA, and discipline (*Zoo U* website[25]).

DeRosier et al. (2012) conducted a research for 254 3rd- and 4th-grade students, and their teachers in two schools in central North Carolina to examine whether the collected data within the gameplay could be used to improve the assessment. They found that detailed game logs of socially relevant player behaviour, if combined with external measures of social skills, could result in valid embedded assessments. Following that, they investigated the correlation between the game performance and teachers' assessments of students' social skills. Their findings showed significant correlations between in-game social skills assessments and teachers' assessment (standard psychological assessments) of the same students as well as high level of engagement. Their study supported the use of interactive games for stealth assessment to measure social skills. The study asserted that the online accessibility, engaging nature and cost effectiveness of *Zoo U* made it an appealing option for assessment to inform decisions regarding implementation

25. Also: 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-PLGWhtSht_FNL.pdf.*

of and tracking social interventions by schools, especially for children in need of social skills interventions.

The study showed that students who demonstrated higher social competency when problem-solving *Zoo U*'s virtual scene were significantly more likely to exhibit positive social, behavioural, and academic adjustment, above and beyond demographic influences, while children who performed poorly on *Zoo U* were significantly more likely to experience negative school-based outcomes.

In another study, Craig, DeRosier, & Watanabe (2015) conducted a comparative assessment, using *Zoo U*, to identify similarities and differences between children in the US and Japan across six specific social skills. The study was conducted in the third and fourth graders in the US and Japan as an assessment tool to assist teachers, clinicians, or parents in determining appropriate social skills interventions for children. For this study, 497 3rd- and 4th-grade grade students (270 from US and 227 from Japan) participated. US students received the original version of *Zoo U*; Japanese children received a fully translated Japanese version of the game. Three performance grades (low, average, high) were developed. Logging-in time and locations of every player click event were captured. Scoring in this game was fully automated by a web-based software, thus eliminating administrative-related errors. All children completed the levels individually on separate computers. Teachers and researchers in the US and Japan were given the same instructions to read to children.

Effective communication in *Zoo U* was assessed via menu choices that would clearly and directly convey the player's thoughts and feelings (e.g., tone of voice, clarity, respect). For example, in one task, the bell rang and the hall monitor informed the child to get to class. The child was to communicate with the hall monitor to find out where the class was and get a hall pass to get there. Based on the cultural differences, it was hypothesised that children in the US would demonstrate higher performance on assessment of effective communication than Japanese children. Cooperation was assessed via a child's ability to work with other students to solve a problem, for instance, to catch a parrot flying around the classroom. The players must request for help appropriately and respond honestly when they receive some suggestions to complete the task. Literature suggested that children in the US were more likely to show responses to peers that emphasise individualism and competition. In comparison, their Asian counterparts, who were more likely to

show responses that favour equality and group enhancement[26]. Thus, it was hypothesised that Japanese students would score higher on cooperation in *Zoo U*.

A MANOVA was conducted to explore the cultural differences in *Zoo U* performance across the social skill scores. The results confirmed that Japanese children scored significantly higher than US students in cooperation, and contrary to the hypothesis, there was no significant difference by cultural group in communication skills. In short, the researchers believed that their findings would preliminarily support the potential of GBA methods to provide efficient and valid social skill assessments to students around the world.

** ECHOES*[27]

As a serious game, *ECHOES* was developed to help acquisition and enhancement of communication skills in young children with Autism Spectrum Conditions (ASCs). In *ECHOES*, children interact with an intelligent virtual character, who acts both as a peer and tutor, over 12 learning activities facilitated by a 42-inch multi-touch LCD display with eye-gaze tracking. The learning activities focus on two particular components of social communication which are identified as the most challenging for children with ASCs: joint attention and symbol use. Both the activities and the agent were designed based on principles of best autism practice and input from users.

29 children were recruited from five special units in mainstream primary schools or schools dedicated to providing care for children with ASCs and/or other disabilities across the UK. In order to describe each child's socio-communicative abilities, a curriculum-based coding assessment was designed to measure different behavioural and emotional outcomes in a meaningful context (e.g., classroom). ELAN (professional audio and video annotation software) was used for the coding scheme. Each video was assessed for 16 main behavioural categories. When an instance of a behaviour was identified, the appropriate code was associated with the period of the video containing the behaviour and its frequency.

The research reported that the *ECHOES* evaluation was one of the first major GBA for autism conducted in real-school contexts, across different

26. For example: Domino, G. (1992). Cooperation and competition in Chinese and American children. *Journal of Cross-Cultural Psychology*, 23(4), 456-467.
27. The study is taken from a project by Bernardini et al. (2014).

schools. Although it could only report the preliminary coarse-grained analysis of children's behaviours in relation to the virtual agent with no significant observation of increased social responsiveness or initiations to real-world contexts, it was successful in many respects. There was evidence that some children benefited from the interaction in the *ECHOES* environment, showing increased number of initiation. Teachers were very positive and agreed that observing children interacting with the agent provided them with valuable and unexpected understanding of the individual children's capabilities.

The biggest limitation of this study was that dealing with heterogeneous population of individuals with ASCs at different ages (ranging from 4–14) made it difficult to create an environment equitable for all users, and restricted the scope for consistent cross-participant comparisons. Another issue was user modelling was based on common features and behaviours across all the users, overlooking individual challenges. To address that, more intimate profiles of the children should be constructed by supplementing automatic user modelling techniques with a direct involvement of teachers, parents and children participants (Bernardini et al., 2014).

**ARGs (Alternative Reality Games)*

ARGs are online interactive narrative and puzzle solving, often involving multiple media and game elements to tell a story that may be affected by players' actions and ideas. A series of media, including websites, instant messenger (IM) conversations, text messages, emails, as well as TV and newspaper adverts and telephone calls reveal the narrative to the players while a puppet master steers players in different directions as the game's story unfolds. ARGs are viewed as being heavily built around social networking and collaboration skills (especially in the context of language learning). Therefore, they can provide a useful educational context and platform for enhancing students' collaboration and communication skills while exploring ideas and views with each other, searching for relevant information and engaging in performing tasks. As well as student-student interaction, the game is considered a suitable platform to increase student-teacher and teacher-teacher interaction.

Connolly, Stansfield, and Hainey (2011) discussed the design, development and evaluation of an ARG (Tower of Babel ARG, arg.uws.ac.uk.), aiming to promote the motivations of secondary school students across Europe in learning foreign languages, as a part of a European Commission project which involved 6 project partners, 328 secondary school students and 95 language

teachers from 17 European countries.

The study asserted that the collaborative nature of ARGs was a suitable vehicle for developing collaborative activities within an educational context. An evaluation of the ARG was conducted using an experimental design of pre-test, ARG intervention, and post-test. In general, students had very positive attitudes towards the ARG, suggesting that the game managed to deliver the motivational experience expected by the students. In addition, besides language learning through using the ARG, students believed that they obtained skills relating to cooperation, collaboration and teamwork.

**Statecraft X*

This fantasy multiplayer strategic game is developed by Chee and his colleagues at the National Institute of Education, Singapore and can be played on iPod or iPhone. The game was designed to assimilate principles and concepts of governance and citizenship, and then examined in in a Social Studies curricula in two Secondary 3 classes (Chee et al., 2010[28]). The study intended to examine students' thinking, beliefs, attitudes as well as their creativity and teamwork after playing *Statecraft X* (on their iPhones) in their Social Studies lessons on governance. In the game, players acted as governors controlling the growth of a town while competing with other fellow students. They should ensure that their citizens were happy and well taken care of, fulfilling the citizens' basic and cultural needs as well as maintaining economic stability and multicultural harmony. It involved three phases: 1) understanding the game and basic governance; 2) advanced development; and 3) expanding the player's sphere of influence by conveyed four key themes: leadership, anticipation of change and staying relevant, reward for work, and work for reward, and a stake for everyone, and opportunities for all. Pre- and post-tests were used to check for experimental effects. Students' understanding of governance and citizenship was assessed through their essays that they submitted at the end of the programme.

The results showed that the experimental class students outperformed the control class as their essays included stronger sense of personal voice, more awareness about local and global issues, and a sense of agency to act with changes. The findings indicated the efficacy of the *Statecraft X* game in curriculum and assessment. The study highlighted an important factor which

28. A similar study: Chee, Y.S., Gwee, S., & Tan, E. M. (2013). Learning to become citizens by enacting governorship in the Statecraft curriculum: An evaluation of learning outcomes. In *Design, Utilization, and Analysis of Simulations and Game-Based Educational Worlds* (pp. 68-94). IGI Global.

was statistically significant in facilitating the learning process and that was the role of non-authoritarian teachers and dialogical pedagogical methods during the intervention. The game was not designed to be used alone, but rather to be incorporated into a dialogic classroom. Then students would learn about the values and dispositions of governance by becoming a virtual governor.

The researchers concluded that such study should be replicated in more classes of Social Studies before generalizing the effects of *Statecraft X*, and overcoming some cultural and pedagogical limitations is necessary in order for GBA to be successfully integrated into Singaporean classrooms.

**** StartUp (http://startup-eu.net)***

As a European project, it was designed to encourage secondary school students' creativity and entrepreneurial skills in the process of replicating a new startup high tech company. The project ran a competition across Europe in 2013 for school teams (3–4 students per team, between 14–18 years old). A set of educational mini-games were developed on a Web2.0 platform, with scenarios encompassing sparkling creativity, building company team, understanding and communicating with clients, marketing the product, and how to develop IT products. Assessment in each mini-game was through the use of quests and how well a player (individually or in a team) performed. Players could play more than once to increase their score (consequently to increase their learning). For example, in the mini-game 'Understanding your clients', the player would play the role of a salesman. By driving a car through part of a city, they had to find potential clients and match them to a suitable product through a number of questions and interpretation of client's answers. The final score would be displayed at the end of the level based on the client-product pairing (more details in Protopsaltis et al., 2013).

**** A case study on assessing collaborative problem-solving in Finnish pre-service teacher education***

In this case study, Ahonen, Häkkinen, & Pöysä-Tarhonen (2018) adopted tasks from the ATC21S assessment portal in the context of pre-service teacher education (24 teacher-students) in Finland to examine their teamwork and collaborative problem-solving. A set of questions was created to measure dispositions towards collaboration, negotiation and collaborative problem-solving, scored on a 7-point Likert-type scale. Then, the ATC21S assessment portal was used to assess participants' skills through CPS game activities

such as *Laughing Clowns, Plant Growth, Balance, Olive oil, and Game of 20*[29]. Each pair of students completed one bundle of five assessment tasks over a period of 90 min. These game-like tasks, mainly in the science and math domains, are related both to curriculum content and to generic skills.

Like in the ATC21S portal, participants' completion of the assessment data was recorded in a log file in an unobtrusive way. The generated data were captured in a process stream data file, and patterns in these data were automatically coded as indicators of the CPS components. Once the indicators were developed, algorithms were programmed to capture the specific sequence of events in the log file data. Each targeted component (e.g., communication, cognitive, etc.) could thus be scaled based on the actions taken by the students together with the online chat discussions during CPS task performance. All actions and chat messages were recorded and time-stamped. The scoring engine automatically coded and scored data to produce reports for both teachers and students.

The findings of this study showed that the level of collaborative problem-solving skills among the experimental group was high. Their social skills were measured higher, as compared to their cognitive skills. Social skills were also connected positively with collaboration and teamwork dispositions—in particular with negotiation. However, the cognitive skills scores did not correlate with teamwork and collaboration dispositions. This indicated that the social aspect of collaborative problem-solving was probably the key for success in these kinds of shared tasks.

Using this web-based portal to measure collaborative problem-solving of pre-service teachers was the first pilot study in employing innovative assessment of complex skills in this context. The study concluded that technology-supported assessment tasks are well-suited to pre-service teacher training; thereby, proving that pre-service teachers are powerful and essential means of sparkling educational changes and adapting to the new learning culture.

29. In developing tasks such as *Laughing Clowns,* test developers mapped the CPS with the major social and cognitive skills dimensions, including Participation, Perspective Taking, Social Regulation, Task Regulation, and Learning and Knowledge Building (Griffin & Care, 2015).

References

Aesaert, K., Van Nijlen, D., Vanderlinde, R., & van Braak, J. (2014). Direct measures of digital information processing and communication skills in primary education: using item response theory for the development and validation of an ICT competence scale. *Computers & Education*, 76, 168–181.

Ahonen, A. K., Häkkinen, P., & Pöysä-Tarhonen, J. (2018). Collaborative problem solving in Finnish pre-service teacher education: A case study. In E. Care, P. Griffin, & B. McGaw (Eds.), *Assessment and Teaching of 21st Century Skills* (pp. 119–130). New York, NY: Springer.

Allen, A., Seeney, M., Boyle, I., & Hancock, F. (2009). The implementation of team based assessment in serious games. In *Proceedings of the 1st Conference in Games and Virtual Worlds for Serious Applications (VS-GAMES '09)*, (pp. 28–35), Coventry, UK.

Almond, R. G., Steinberg, L. S., & Mislevy, R. J. (2003). *A four-process architecture for assessment delivery, with connections to assessment design* (Vol. 616). Los Angeles: University of California Los Angeles Centre for Research on Evaluations, Standards and Student Testing (CRESST).

Ananiadou, K., & Claro, M. (2009). *21st century skills and competences for new millennium learners in OECD countries*. OECD Education Working Papers, No. 41. Paris: OECD Publishing.

Armstrong, A., & Georgas, H. (2006). Using interactive technology to teach information literacy concepts to undergraduate students. *Reference Services Review*, 34(4), 491–497.

ATC21S – Assessment & Teaching of 21st century skills. (2009). *Transforming education: assessing and teaching 21st century skills* [Assessment Call to Action]. Retrieve from http://atc21s.org/wp-content/uploads/2011/04/Cisco-Intel-Microsoft-Assessment-Call-to-Action.pdf

Bachen, C. M., Hernández-Ramos, P. F., & Raphael, C. (2012). Simulating REAL LIVES: Promoting global empathy and interest in learning through simulation games. *Simulation & Gaming*, 43(4), 437–460.

Bagley, E., & Shaffer, D. W. (2009). When people get in the way: Promoting civic thinking through epistemic gameplay. *International Journal of Gaming and Computer-Mediated Simulations (IJGCMS)*, 1(1), 36–52.

Baker, E. L., Chung, G. K. W. K., & Delacruz, G. C. (2012). The best and future uses of assessment in games. In M. Mayrath, J. Clarke-Midura, & D. H. Robinson (Eds.), *Technology-based assessments for 21st century skills: Theoretical and practical implications from modern research* (pp. 248–299). Charlotte, NC: Information Age.

Barab, S. A., Sadler, T. D., Heiselt, C., Hickey, D., & Zuiker, S. (2007). Relating narrative, inquiry, and inscriptions: Supporting consequential play. *Journal of science education and technology*, 16(1), 59–82.

Barbera, E. (2009). Mutual feedback in e-portfolio assessment: an approach to the netfolio system. *British journal of educational technology*, 40(2), 342–357.

Bauer, M., Wylie, C., Jackson, T., Mislevy, B., Hoffman-John, E., John, M., & Corrigan,

S. (2017). Why Video Games can be a Good Fit for Formative Assessment. *Journal of Applied Testing Technology*, 18(S1), 19–31.

Bellotti, F., Kapralos, B., Lee, K., Moreno-Ger, P., & Berta, R. (2013). Assessment in and of serious games: An overview. *Advances in Human-Computer Interaction*, 2013, 1–11.

Bennett, R. E. (2011). Formative assessment: A critical review. *Assessment in Education: Principles, Policy & Practice*, 18(1), 5–25.

Bente, G., & Breuer, J. (2009). Making the implicit explicit. In U. Ritterfeld, M. Cody & P. Vorderer (Eds.), *Serious games: Mechanisms and effects* (pp. 322–343). New York, NY: Routledge.

Bernardini, S., Porayska-Pomsta, K., & Smith, T. J. (2014). ECHOES: An intelligent serious game for fostering social communication in children with autism. *Information Sciences*, 264, 41–60.

Binkley, M., Erstad, O., Herman, J., Raizen, S., & Ripley, M. (2012). Defining 21st century skills. In P. Griffin, B. McGaw, & E. Care (Eds.), *Assessment and teaching of 21st century skills* (pp. 17–66). Dordrecht, Germany: Springer.

Bowe, R., Ball, S. J., & Gold, A. (2017). *Reforming education and changing schools: Case studies in policy sociology*. London: Routledge.

Bridgeman, B., Lennon, M. L., & Jackenthal, A. (2003). Effects of screen size, screen resolution, and display rate on computer-based test performance. *Applied Measurement in Education*, 16(3), 191–205.

Bujanda, M. E., Muñoz, L., & Zúñiga, M. (2018). Initiatives and Implementation of Twenty-First Century Skills Teaching and Assessment in Costa Rica. In *Assessment and Teaching of 21st Century Skills* (pp. 163–178). New York, NY: Springer.

Care, E., Griffin, P., & McGaw, B. (2018). *Assessment and Teaching of 21st Century Skills*. New York, NY: Springer.

Care, E., & Kim, H. (2018). Assessment of Twenty-first century skills: The issue of authenticity. In *Assessment and Teaching of 21st Century Skills* (pp. 21–39). New York, NY: Springer.

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

Carson, S., Peterson, J.B & Higgins, D. M. (2005). Reliability, validity and factor structure of the Creative Achievement Questionnaire. *Creativity Research Journal*, 17(1), 37–50.

CASEL - Collaborative for Academic, Social, and Emotional Learning. (2019). Assessment. Retrieved from https://casel.org/assessment-references

CEDEFOP - European Centre for the Development of Vocational Training. (2008). *Terminology of European Education and Training Policy*. Retrieved from http://www.cedefop.europa.eu/files/4064_en.pdf

Chang, M. (2010). Web-based Multiplayer Online Role Playing Game (MORPG) for Assessing Students' Java Programming Knowledge and Skills. In *Proceedings of 2010 IEEE International Conference on Digital Game and Intelligent Toy Enhanced Learning*.

Chang, C. C., & Tseng, K. H. (2009). Use and performances of web-based portfolio assessment. *British Journal of Educational Technology*, 40(2), 358–370.

Chang, K. E., Wu, L. J., Weng, S. E., & Sung, Y. T. (2012). Embedding game-based problem-solving phase into problem-posing system for mathematics learning. *Computers & Education*, 58(2), 775–786.

Chee, Y. S., Mehrotra, S., & Ong, J. C. (2014). Facilitating dialog in the game-based learning classroom: Teacher challenges reconstructing professional identity. *Digital Culture & Education*, 6(4), 299–316.

Chee, Y.S., Tan, E. M., & Liu, Q. (2010, April). Statecraft X: Enacting citizenship education using a mobile learning game played on Apple iPhones. In *Wireless, Mobile and Ubiquitous Technologies in Education (WMUTE), 2010 6th IEEE International Conference* on (pp. 222–224). IEEE.

Chee, Y.S., Tan, K. C. D., Tan, E. M., & Jan, M. (2012). Learning chemistry performatively: Epistemological and pedagogical bases of design-for-learning with computer and video games. In K. C. D. Tan, & M. Kim, M. (Eds.), *Issues and Challenges in Science Education Research* (pp. 245–262). Dordrecht, Germany: Springer

Chen, S., & Michael, D. (2005). *Proof of Learning: Assessment in Serious Games*. Retrieved from http://ww.w.cedma-europe.org/newsletter%20articles/misc/Proof%20of%20Learning%20-%20Assessment%20in%20Serious%20games%20(Oct%2005).pdf

Chin, J., Dukes R., & Gamson, W. (2009). Assessment in simulation and gaming: A review of the last 40 years. *Simulation & Gaming*, 40(4), 553–568.

Christian, M. S., Edwards, B. D., & Bradley, J. C. (2010). Situational judgment tests: Constructs assessed and a meta-analysis of their criterion-related validities. *Personnel Psychology*, 63(1), 83–117.

Chu, C. B. L., Yeung, A. H. W., & Chu, S. K. W. (2012). Assessment of students' information literacy: A case study of a secondary school in Hong Kong. Paper presented at CITE Research Symposium 2012. Hong Kong: The University of Hong Kong.

Chu, S. K. W. (2012). Assessing Information Literacy: A Case Study of Primary 5 Students in Hong Kong. *School Library Research*, 15, 1–24.

Chung, G. K., & Baker, E. L. (2003). An exploratory study to examine the feasibility of measuring problem-solving processes using a click-through interface. *The Journal of Technology, Learning and Assessment*, 2(2). Retrieved from https://ejournals.bc.edu/ojs/index.php/jtla/article/view/1662

Claro, M., Preiss, D. D., San MartíN, E., Jara, I., Hinostroza, J. E., Valenzuela, S., & Nussbaum, M. (2012). Assessment of 21st century ICT skills in Chile: Test design and results from high school level students. *Computers & Education*, 59(3), 1042–1053.

Comfort, K. B., & Timms, M. (2018). A Twenty-First Century Skills Lens on the Common Core State Standards and the Next Generation Science Standards. In E. Care, P. Griffin, & B. McGaw (Eds.), *Assessment and Teaching of 21st Century Skills* (pp.

131–144). New York, NY: Springer.
Connolly, T. M., Boyle, E. A., MacArthur, E., Hainey, T., & Boyle, J. M. (2012). A systematic literature review of empirical evidence on computer games and serious games. *Computers & Education*, 59(2), 661–686.
Connolly, T. M., Stansfield, M., & Hainey, T. (2011). An alternate reality game for language learning: ARGuing for multilingual motivation. *Computers & Education*, 57(1), 1389–1415.
Cowley, B., Charles, D., Black, M. and Hickey, R. (2008). Toward an understanding of flow in video games. *Computers in Entertainment*, 6(2), 1–28.
Craig, A. B., DeRosier, M. E., & Watanabe, Y. (2015). Differences between Japanese and US children's performance on “Zoo U”: A game-based social skills assessment. *Games for Health Journal*, 4(4), 285–294.
Csapó, B., Ainley, J., Bennett, R. E., Latour, T., & Law, N. (2012). Technological issues for computer-based assessment. In P. Griffin, B. McGaw, & E. Care (Eds.), *Assessment and Teaching of 21st Century Skills* (pp. 143–230). Dordrecht, Germany: Springer.
Csapó, B., Molnár, G., & Tóth, K. R. (2009). Comparing Paper-and-Pencil and Online Assessment of Reasoning Skills. In F. Scheuermann & J. Björnsson (Eds.), *The Transition to Computer-based Assessment* (pp. 120–125). Luxembourg: Office for official publications of the European communities.
Csikszentmihalyi, M. (1990) *Flow: The Psychology of Optimal Experience*. New York, NY: Harper & Row.
Darling-Hammond, L. (2012). Policy frameworks for new assessments. In P. Griffin, B. McGaw, & E. Care (Eds.), *Assessment and Teaching of 21st Century Skills* (pp. 301–339). Dordrecht, Germany: Springer.
Darling-Hammond, L., Adamson, F., & Abedi, J. (2010). *Beyond basic skills: The role of performance assessment in achieving 21st century standards of learning*. Stanford, UK: Stanford Centre for Opportunity Policy in Education.
Davis, M. H. (1983). Measuring individual differences in empathy: Evidence for a multidimensional approach. *Journal of Personality and Social Psychology*, 44(1), 113–126.
Dede, C. (2010). Comparing frameworks for 21st century skills. *21st Century Skills: Rethinking How Students Learn*, 20, 51–76.
Delacruz, G. C. (2011). Games as Formative Assessment Environments: Examining the Impact of Explanations of Scoring and Incentives on Math Learning, Game Performance, and Help Seeking. CRESST Report 796. *National Centre for Research on Evaluation, Standards, and Student Testing (CRESST)*.
DeRosier, M. E., Craig, A. B., & Sanchez, R. P. (2012). Zoo U: A stealth approach to social skills assessment in schools. *Advances in Human-Computer Interaction*. Retrieved from http://dx.doi.org/10.1155/2012/654791
Deterding, S., Sicart, M., Nacke, L., O’Hara, K., & Dixon, D. (2011, May). Gamification: Toward a definition. In *Proceedings of the CHI 2011 Gamification Workshop* (pp. 2425–2428). Vancouver, Canada.

DiCerbo, K. E. (2017). Building the evidentiary argument in game-based assessment. *Journal of Applied Testing Technology*, 18(S1), 7–18.

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

Earl, L. M. (2012). *Assessment as learning: Using classroom assessment to maximize student learning*. Thousand Oaks, CA: Corwin Press.

Ecclestone, K. (2010). *Transforming formative assessment in lifelong learning*. London: McGraw-Hill Education.

Eid, M. E., & Diener, E. E. (2006). *Handbook of Multimethod Measurement in Psychology*. Washington, DC: American Psychological Association.

Eseryel, D., Ifenthaler, D., & Ge, X. (2013). Validation study of a method for assessing complex ill-structured problem solving by using causal representations. *Educational Technology Research and Development*, 61(3), 443–463.

Fan, L., Quek, K. S., Koay, P. L., Ng, J., Pereira-Mendoza, L., Yeo, S. M., & Zhu, Y. (2008). *Integrating new assessment strategies into mathematics classrooms: An exploratory study in Singapore primary and secondary schools* (Final Research Report). Singapore: National Institute of Education, Centre for Research in Pedagogy and Practice. Retrieved from http://crpp.nie.edu.sg/~pubs/CRP24_03FLH_FinalResRpt.pdf

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

Feng, M., Heffernan, N. T., & Beck, J. E. (2009, July). Using Learning Decomposition to Analyse Instructional Effectiveness in the ASSISTment System. In Dimitrova, V., Mizoguchi, R., du Boulay, B. & Graesser, A. (Eds.), *Proceedings of the 14th International Conference on Artificial Intelligence in Education* (pp. 523–530). Brighton, UK.

Fletcher, G. (2007, October). Assessing learning from a holistic approach: Creating a balanced system of learning assessment. Paper presented the *Congreso Internacional Evaluacion Factor de Calidad Educativa*, Queretaro, Mexico.

Fraillon, J., Schulz, W., Friedman, T., Ainley, J., & Gebhardt, E. (2015). International Computer and Information Literacy Study: *ICILS 2013*. Technical Report, Amsterdam.

Fu, F. L., Su, R. C., & Yu, S. C. (2009). EGameFlow: A scale to measure learners' enjoyment of e-learning games. *Computers & Education*, 52(1), 101–112.

Funke, J., Fischer, A., & Holt, D. V. (2018). Competencies for complexity: problem solving in the twenty-first century. In E. Care, P. Griffin, & B. McGaw (Eds.), *Assessment and Teaching of 21st Century Skills* (pp. 41–53). New York, NY: Springer.

Garrett, N., Thoms, Alrushiedat, N., & Ryan, T. (2009). Social ePortfolios as the new course management system. *On the Horizon*, 17(3), 197–207.

Gee, J. P. (2009). Deep Learning Properties of Good Video Games: How Far Can They Go? In U. Ritterfield, M. Cody, & P. Vorderer (Eds.), *Serious Games: Mechanisms and Effects* (pp. 67–82). New York, NY: Routledge.

Gee, J. P. (2004). *Situated Language and Learning: A Critique of Traditional Schooling*. London: Routledge.

Gentile, D. A., Anderson, C. A., Yukawa, S., Ihori, N., Saleem, M., Ming, L. K., & Rowell Huesmann, L. (2009). The effects of prosocial video games on prosocial behaviors: International evidence from correlational, longitudinal, and experimental studies. *Personality and Social Psychology Bulletin*, 35(6), 752–763.

Gipps, C. V. (1994). *Beyond Testing: Towards a Theory of Educational Assessment*. London, UK: Falmer Press.

Girard, C., Ecalle, J., & Magnan, A. (2013). Serious games as new educational tools: how effective are they? A meta-analysis of recent studies. *Journal of Computer Assisted Learning*, 29(3), 207–219.

Gobert, J. D., Kim, Y. J., Sao Pedro, M. A., Kennedy, M., & Betts, C. G. (2015). Using educational data mining to assess students' skills at designing and conducting experiments within a complex systems micro-world. *Thinking Skills and Creativity*, 18, 81–90.

Gordon, J., Halsz, G., Krawczyk, M., Leney, T., Michel, A., Pepper, D., Putkiewicz, E., and Wisniewski, W. (2009). Key competences in Europe. Opening doors for lifelong learners across the school curriculum and teacher education. *Case-Center for Social and Economic Research*, Warsaw.

Graesser, A., Britt, A., Millis, K., Wallace, P., Halpern, D., Cai, Z., & Forsyth, C. (2010). Critiquing media reports with flawed scientific findings: Operation ARIES! A game with animated agents and natural language trialogues. In *International Conference on Intelligent Tutoring Systems* (pp. 327–329). Berlin: Springer.

Graesser A., Dowell N., Clewley D. (2017) Assessing Collaborative Problem Solving Through Conversational Agents. In: von Davier A., Zhu M., Kyllonen P. (Eds.), *Innovative Assessment of Collaboration. Methodology of Educational Measurement and Assessment* (pp. 65–80). New York: Springer.

Greenstein, L. M. (2012). *Assessing 21st Century Skills: A Guide to Evaluating Mastery and Authentic Learning*. Thousand Oaks, CA: Sage.

Greiff, S., & Funke, J. (2009). *Measuring Complex Problem Solving: The MicroDYN Approach*. In F. Scheuermann & J. Bjjörnsson (Eds.), *The transition to computer-based assessment—Lessons learned from largescale surveys and implications for testing* (pp. 157–163). Luxembourg: European Commission Joint Research Centre.

Griffin, P., & Care, E. (Eds.). (2015). *Assessment and Teaching of 21st Century Skills: Methods and Approach*. Dordrecht, Germany: Springer.

Griffin, P., Care, E., & McGaw, B. (2012). *Assessment and Teaching of 21st Century Skills*. Dordrecht, Germany: Springer.

Griffin, P. J., & Nix, P. (1991). *Educational Assessment and Reporting: A New Approach*. London, UK: Harcourt Brace Jovanovich.

Hainey, T., Connolly, T. M., Chaudy, Y., Boyle, E., Beeby, R., & Soflano, M. (2015).

Assessment integration in serious games. In Information Resources Management Association, *Gamification: Concepts, methodologies, tools, and applications* (pp. 515–540). Hershey, PA: IGI Global.

Haldane, S. (2009). Delivery platforms for national and international computer-based surveys. In F. Scheuermann, & J. Bjjörnsson (Eds.) *The transition to computer-based assessment: New approaches to skills assessment and implications for large-scale testing* (pp. 63-67). Luxembourg: European Commission Joint Research Centre.

Halpern, D. F. (1998). Teaching critical thinking for transfer across domains: Disposition, skills, structure training and metacognitive monitoring. *American Psychologist*, 53(4), 449–455.

He, J., & Van De Vijver, F. J. (2015). Effects of a general response style on cross-cultural comparisons: Evidence from the Teaching and Learning International Survey. *Public Opinion Quarterly*, 79(S1), 267–290.

Heggestad, E. D., Morrison, M., Reeve, C. L., & McCloy, R. A. (2006). Forced-choice assessments of personality for selection: Evaluating issues of normative assessment and faking resistance. *Journal of Applied Psychology*, 91(1), 9.

Hohlfeld, T. N., Ritzhaupt, A. D., & Barron, A. E. (2010). Development and validation of the Student Tool for Technology Literacy (ST2L). *Journal of Research on Technology in Education*, 42(4), 361–389.

Holmes, L. M. (2012). *The effects of project-based learning on 21st century skills and no child left behind accountability standards*. Doctoral dissertation, University of Florida.

Inal, Y., & Cagiltay, K. (2007). Flow experiences of children in an interactive social game environment. *British Journal of Educational Technology*, 38(3), 455–464.

Johnson, D. W., Johnson, R. T., & Holubec, E. J. (1994). *The new circles of learning: Cooperation in the classroom and school*. Alexandria, VA: Association for Supervision and Curriculum Development (ASCD).

Kato, P. M., & de Klerk, S. (2017). Serious games for assessment: welcome to the jungle. *Journal of Applied Testing Technology*, 18(S1), 1–6.

Katz, I. R., & Macklin, A. S. (2007). Information and communication technology (ICT) literacy: Integration and assessment in higher education. *Systemics, Cybernetics and Informatics*, 5(4), 50–55.

Kim, Y. J., & Shute, V. J. (2015). The interplay of game elements with psychometric qualities, learning, and enjoyment in game-based assessment. *Computers & Education*, 87, 340–356.

King, G., & Wand, J. (2007). Comparing incomparable survey responses: Evaluating and selecting anchoring vignettes. *Political Analysis*, 15(1), 46–66.

Koenig, Judith Anderson (Ed). (2011). *Assessing 21st Century Skills: Summary of a workshop*. Washington D.C.: National Academies Press.

Koh, K. H., Tan, C., & Ng, P. T. (2012). Creating thinking schools through authentic assessment: The case in Singapore. *Educational Assessment, Evaluation and Accountability*, 24(2), 135–149.

Ku, K. Y. (2009). Assessing students' critical thinking performance: Urging for measurements using multi-response format. *Thinking Skills and Creativity*, 4(1), 70–76.

Kuncel, N. R. (2011, January). *Measurement and Meaning of Critical Thinking*. Report presented at the National Research Council's 21st Century Skills Workshop, Irvine, CA.

Kyllonen, P. C. (2012). Measurement of 21st century skills within the common core state standards. In *Invitational Research Symposium on Technology Enhanced Assessments*, (pp. 7–8). https://www.ets.org/Media/Research/pdf/session5-kyllonen-paper-tea2012.pdf.

Lai, E. R., & Viering, M. (2012). *Assessing 21st Century Skills: Integrating Research Findings*. New York: Pearson.

Law, N., Yuen, H. K., Shum, M., & Lee, Y. (2007). *Phase (II) study on evaluating the effectiveness of the 'empowering learning and teaching with information technology' strategy (2004/2007). Final report*. Hong Kong: Hong Kong Education Bureau.

Lederman, N. G., & Abell, S. K. (Eds.). (2014). *Handbook of Research on Science Education* (Vol. 2). London: Routledge.

Lipnevich, A. A., MacCann, C., & Roberts, R. D. (2013). Assessing non-cognitive constructs in education: A review of traditional and innovative approaches. In D.H. Saklofske, V.L. Schwean, & C.R. Reynolds, (Eds.), *The Oxford Handbook of Child Psychological Assessment*. Oxford: Oxford University Press.

DOI: 10.1093/oxfordhb/9780199796304.013.0033

Loader, B. D. (Ed.). (2007). *Young citizens in the digital age: Political engagement, young people and new media*. London: Routledge.

Loh, C. S., Anantachai, A., Byun, J., & Lenox, J. (2007, July). Assessing what players learned in serious games: In situ data collection, information trails, and quantitative analysis. In *10th International Conference on Computer Games*: *AI, Animation, Mobile, Educational & Serious Games*, (pp. 25–28), Louisville, KY.

MacCann, C., Lievens, F., Libbrecht, N., & Roberts, R. D. (2016). Differences between multimedia and text-based assessments of emotion management: An exploration with the multimedia emotion management assessment (MEMA). *Cognition and Emotion*, 30(7), 1317–1331.

Malone, T. W., & Lepper, M. R. (1987). Making learning fun: A taxonomy of intrinsic motivations for learning. In R. E. Snow, & Marshall J. Farr, (Eds.), *Aptitude, Learning and Instruction*, Vol. 3 (pp. 223–253). Hillsdale, NJ: Lawrence Erlbaum.

Marczewski, A. (2013). What's the difference between Gamification and Serious Games. *Andrzej's Blog*. Retrieved from https://www.gamasutra.com/blogs/AndrzejMarczewski/20130311/188218/Whats_the_difference_between_Gamification_and_Serious_Games.php

Mcdaniel, M. A., Hartman, N. S., Whetzel, D. L., & GRUBB III, W. L. (2007). Situational judgment tests, response instructions, and validity: A meta-analysis. *Personnel Psychology*, 60(1), 63–91.

Ministry of Education, Singapore. (2010). *Elaboration of the MOE 21st century*

competencies. Retrieved from https://www.moe.gov.sg/education/education-system/21st-century-competencies

Mislevy, R. (1994). Evidence and inference in educational assessment. *Psychometrika*, 59(4), 439–483.

Mislevy, R. J., Behrens, J. T., Dicerbo, K. E., Frezzo, D. C., & West, P. (2012). Three things game designers need to know about assessment. In D. Ifenthaler, D Eseryel, X Ge, (Eds.), *Assessment in Game-based Learning* (pp. 59–81). New York, NY: Springer.

Mislevy, R. J., Oranje, A., Bauer, M. I., von Davier, A., Hao, J., Corrigan, S., & John, M. (2014). *Psychometric Considerations in Game-based Assessment*. New York, NY: Institute of Play.

Montgomery, K. (2001). *Authentic Assessment: A guide for elementary teachers*. New York: Longman.

National Research Council (NRC). (2007). *Taking Science to School: Learning and teaching science in grades K-8*. Washington D.C.: National Academies Press.

Ng, P. T. (2004). Students' perception of change in the Singapore education system. *Educational Research for Policy and Practice*, 3(1), 77–92.

Nisbett, R. E., & Wilson, T. D. (1977). The halo effect: Evidence for unconscious alteration of judgments. *Journal of Personality and Social Psychology*, 35(4), 250–256.

Nitko, A. J. (2001). *Educational Assessment of Students* (3rd ed.). Upper Saddle River, NJ: Merrill.

OECD. (2013). *Draft PISA Collaborative Problem Solving Framework*. Paris, France: OECD Publishing. Retrieved from http://www.oecd.org/pisa/pisaproducts/

OECD. (2005). *The definition and selection of key competencies executive summary*. Retrieved from http://www.oecd.org/pisa/35070367.pdf

O'Neil, H. F., & Chuang, S. H. (2008). Measuring collaborative problem solving in low-stakes tests. In E. L. Baker, J. Dickieson, W. Wulfeck & H. F. O'Neil (Eds.), *Assessment of Problem Solving Using Simulations* (pp. 177–199). Mahwah, NJ: Lawrence Erlbaum Associates.

O'Neill, H. F., Chuang, S. H., & Chung, G. K. (2003). Issues in the computer-based assessment of collaborative problem solving. *Assessment in Education: Principles, Policy & Practice*, 10(3), 361–373.

O'Neill, H. F., Wainess, R., & Baker, E. L. (2005). Classification of learning outcomes: evidence from the computer games literature. *The Curriculum Journal*, 16(4), 455–474.

P21 - Partnership for 21st Century Skills. (2009). *Framework for 21st Century Learning*. Retrieved from http://www.p21.org/documents/P21Framework.pdf.

P21 - Partnership for 21st Century Skills. (2009). *Framework Definitions*. Retrieved from http://www.p21.org/documents/P21_Framework_Definitions.pdf.

P21 - Partnership for 21st Century Skills. (2009). *Assessment: A 21st Century Skills Implementation Guide*. Retrieved from http://p21.org/documents/p21-stateimp_assessment.pdf.

Pead, D. (2012). World class tests: Summative assessment of problem-solving using technology. *Educational Designer*, 2(5). Retrieved from http://www.educationaldesigner.org//ed/volume2/issue5/article18/.

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

Pepper, D. (2011). Assessing key competences across the curriculum—and Europe. *European Journal of Education*, 46(3), 335–353.

Pepper, D. (2013). KeyCoNet 2013 literature review: Assessment for key competencies. *European Schoolnet and Key Competence Network on School Education (KeyCoNet). Brussels: European Commission*.

Petko, D., Prasse, D., & Cantieni, A. (2018). The interplay of school readiness and teacher readiness for educational technology integration: A structural equation model. *Computers in the Schools*, 35(1), 1–18.

Plichart, P., Jadoul, R., Vandenabeele, L., & Latour, T. (2004). TAO, a collaborative distributed computer-based assessment framework built on Semantic Web standards. In *Communication at the Advances in Intelligent Systems–Theory and Application AISTA2004*, Luxembourg.

Plomp, T., Anderson, R. E., Law, N., & Quale, A. (2009). *CrossNational Information and Communication Technology Policies and Practices in Education* (2nd ed.). IAP.

Poon, C. L., Tan, S., Cheah, H. M., Lim, P. Y., & Ng, H. L. (2015). Student and teacher responses to collaborative problem solving and learning through digital networks in Singapore. In *Assessment and Teaching of 21st Century Skills* (pp. 199–212). Dordrecht, Germany: Springer.

Prelec, D. (2004). A Bayesian truth serum for subjective data. *Science*, 306(5695), 462–466.

Prensky, M. (2001). Digital natives, digital immigrants part 1. *On The Horizon*, 9(5), 1–6.

Programme for the International Assessment of Adult Competencies (PIAAC). (2018). *Survey of Adult Skills*. Retrieved from http://www.oecd.org/skills/piaac/

Protopsaltis, A., Hainey, T., Borosis, S., Connolly, T., Copado, J., & Hezner, S. (2013). Startup_EU: Using game-based learning and web 2.0 technologies to teach entrepreneurship to secondary education students. In *7th European Conference on Games Based Learning*.

Quellmalz, E. S., & Pellegrino, J. W. (2009). Technology and testing. *Science*, 323(5910), 75–79.

Ramalingam, D., & Adams, R. J. (2018). How Can the Use of Data from Computer-Delivered Assessments Improve the Measurement of Twenty-First Century Skills? In E. Care, P. Griffin, B. McGaw, (Eds.), *Assessment and Teaching of 21st Century Skills*, (pp. 225–238). Springer, Cham.

Ramos, G., & Schleicher, A. (2016). *Global Competency for An Inclusive World*. France: OECD.

RAND & Asia Society, 2013, *Measuring 21st Century Competencies: Guidance for*

educators: http://asiasociety.org/files/gcenmeasuring21cskills.pdf.
Raven, J. C., & John Hugh Court. (1998). *Raven's Progressive Matrices and Vocabulary Scales*. Oxford: Oxford Pyschologists Press.
Redecker, C. (2013). The use of ICT for the assessment of key competences. *Joint Research Centre of the European Commission Scientific and Policy Report*. Luxembourg: Joint Research Centre of the European Commission.
Resnick, L. B., Spillane, J. P., Goldman, P., & Rangel, E. S. (2010). Implementing innovation: from visionary models to everyday practice. In H. Dumont, D. Istance, & F. Benavides Francisco, (Eds.), *The Nature of Learning: Using research to inspire practice* (pp. 285-315). Paris: OECD Publishing.
Rosen, Y., & Foltz, P. W. (2014). Assessing collaborative problem solving through automated technologies. *Research & Practice in Technology Enhanced Learning*, 9(3), 389–410.
Rosen, Y., & Tager, M. (2013). *Computer-based assessment of collaborative problem-solving skills: Human-to-agent versus human-to-human approach*. Boston, MA: Pearson Education.
Roskos, K., & Neuman, S. B. (2012). Formative assessment: Simply, no additives. *The Reading Teacher*, 65(8), 534–538.
Rupp, A.A., Gushta, M., Mislevy, R.J., & Shaffer, D.W. (2010). Evidence-centred design of epistemic games: Measurement principles for complex learning environments. *Journal of Technology, Learning, and Assessment*, 8(4). Retrieved from http://www.jtla.org
Sackett, P. R. (2006). Faking and coaching effects on non-cognitive predictors. Paper presented at the *ETS Mini-conference on Faking in Non-Cognitive Assessments*. Princeton, NJ: Educational Testing Service.
Saxton, E., Belanger, S., & Becker, W. (2012). The critical thinking analytic rubric (CTAR): Investigating intra-rater and inter-rater reliability of a scoring mechanism for critical thinking performance assessment. *Assessing Writing*, 17(4), 251–270.
Scalise, K., & Gifford, B. (2006). Computer-based assessment in e-learning: A framework for constructing intermediate constraint questions and tasks for technology platforms. *The Journal of Technology, Learning and Assessment*, 4(6). Retrieved from http://www.jtla.org
Scardamalia, M., Bransford, J., Kozma, B., & Quellmalz, E. (2012). New assessments and environments for knowledge building. In E. Care, P. Griffin, B. McGaw, (Eds.), *Assessment and Teaching of 21st Century Skills*, (pp. 231–300). Springer.
Scheuermann, F., & Björnsson, J. (2009). The transition to computer-based assessment. *Luxembourg: Office for official publications of the European communities*.
Shaffer, D. W., Collier, W., & Ruis, A. R. (2016). A tutorial on epistemic network analysis: Analyzing the structure of connections in cognitive, social, and interaction data. *Journal of Learning Analytics*, 3(3), 9–45.
Shaffer, D. W., & Gee, J.P. (2007). Epistemic games as education for innovation. In J. Underwood & J. Dockrell, (Eds.), *Learning Through Digital Technologies*,

(pp. 71–82). Leicester, UK: British Journal of Educational Psychology. http://epistemicgames.org/cv/papers/Ep-games-innovation-Shaffer-Gee-BPS-2007.pdf.

Shaffer, D. W., & Gee, J. P. (2012). The right kind of GATE: Computer games and the future of assessment. In G Schraw, MC Mayrath, J ClarkeMidura, & DH Robinson, (Eds.), *Technology-based Assessments for 21st Century Skills*: Theoretical and practical implications from modern research (pp. 211–228). Charlotte, NC: Information Age Publications.

Sharpe, L., & Gopinathan, S. (2002). After effectiveness: New directions in the Singapore school system? *Journal of Education Policy*, 17(2), 151–166.

Shute, V. J. (2009). Simply assessment. *International Journal of Learning and Media*, 1(2), 1–11.

Shute, V. J., Hansen, E. G., & Almond, R. G. (2008). You can't fatten A hog by weighing It–Or can you? evaluating an assessment for learning system called ACED. *International Journal of Artificial Intelligence in Education*, 18(4), 289–316.

Shute, V., Ke, F., & Wang, L. (2017). Assessment and adaptation in games. In P. Wouters & H. van Oostendorp, (Eds.), *Instructional techniques to facilitate learning and motivation of serious games* (pp. 59–78). Springer.

Shute, V. J., Masduki, I., & Donmez, O. (2010). Conceptual framework for modeling, assessing and supporting competencies within game environments. *Technology, Instruction, Cognition & Learning*, 8(2), 137–161.

Shute, V. & Ventura, M. (2013). *Stealth assessment: Measuring and supporting learning in video games*. London: MacArthur Foundation.

Shute, V. J., Ventura, M., Bauer, M., & Zapata-Rivera, D. (2009). Melding the power of serious games and embedded assessment to monitor and foster learning. *Serious games: Mechanisms and effects*, 2, 295–321.

Silvia, P. J., Wigert, B., Reiter-Palmon, R. & Kaufman, J. C. (2012). Assessing creativity with self-reports scales: A review and empirical evaluation. *Psychology of Aesthetics, Creativity and the Arts*, 6(1), 19–34.

Simon, M., & Forgette-Giroux, R. (2000). Impact of a content selection framework on portfolio assessment at the classroom level. *Assessment in Education: Principles, Policy & Practice*, 7(1), 83–100.

Soland, J., Hamilton, L. S., & Stecher, B. M. (2013). *Measuring 21st Century Competencies*. Global Cities Education Network: New York, NY: Asia Society/ RAND Corporation.

Sourmelis, T., Ioannou, A., & Zaphiris, P. (2017). Massively Multiplayer Online Role Playing Games (MMORPGs) and the 21st century skills: A comprehensive research review from 2010 to 2016. *Computers in Human Behavior*, 67, 41–48.

Spector, J. M., & Koszalka, T. A. (2004). *The DEEP methodology for assessing learning in complex domains* (Final report to the National Science Foundation Evaluative Research and Evaluation Capacity Building). Syracuse, NY: Syracuse University.

Squire, K. (2011). *Video Games and Learning: Teaching and Participatory Culture in the Digital Age. Technology, Education--Connections (the TEC Series)*. New York, NY: Teachers College Press.

Stark, S., Chernyshenko, O. S., & Drasgow, F. (2005). An IRT approach to constructing and scoring pairwise preference items involving stimuli on different dimensions: The multi-unidimensional pairwise-preference model. *Applied Psychological Measurement*, 29(3), 184–203.

Susi, T., Johannesson, M., & Backlund, P. (2007). *Serious games: An overview*. Sweden: University of Skövde School of Humanities and Informatics.

Sweetser, P., & Wyeth, P. (2005). GameFlow: a model for evaluating player enjoyment in games. *ACMComputers in Entertainment*, 3(3), 1–24.

Tan, J. P. L., Choo, S. S. L., Kang, T., & Liem, G. A. D. (2017). Educating for twenty-first century competencies and future-ready learners: Research perspectives from Singapore. *Asia Pacific Journal of Education*, 37(4), 425–436.

Timperley, H. (2008). *Teacher Professional Learning and Development*. Paris: UNESCO.

Tisani, N. (2008). Challenges in producing a portfolio for assessment: in search of underpinning educational theories. *Teaching in Higher Education*, 13(5), 549–557.

Torres, R. J. (2009). *Learning on a 21st century platform: Gamestar mechanic as a means to game design and systems-thinking skills within a nodal ecology* (Doctoral dissertation). New York: New York University.

Treffinger, D. J., Young, G. C., Selby, E. C., & Shepardson, C. (2002). Assessing creativity: A guide for educators. University of Connecticut: National Research Centre on the Gifted and Talented.

Trier, U. (2003). Twelve countries contributing to DeSeCo: A summary report. In D. Rychen, L. Salganik, & M. McLaughlin (Eds.), *Definition and selection of key competences. Contributions to the second DeSeCo symposium*, (pp. 7–59). Neuchatel: Swiss Federal Statistical Office.

Trilling, B., & Fadel, C. (2009). *21st Century Skills: Learning for life in our times*. New Jersey: John Wiley & Sons.

Tsai, C. C., Lin, S. S., & Yuan, S. M. (2001). Students' use of web-based concept map testing and strategies for learning. *Journal of Computer Assisted Learning*, 17(1), 72–84.

Turner, J. C. (1995). The influence of classroom contexts on young children's motivation for literacy. *Reading Research Quarterly*, 30(3), 410–441.

UNESCO. (2012). *Competencies*. Education. Retrieved from http://www.unesco.org/new/en/education/themes/strengthening-education-systems/quality-framework/desired-outcomes/competencies/.

Van Horn, R. (2003). Computer adaptive tests and computer-based tests. *Phi Delta Kappan*, 84(8), 567–631.

Vlug, K. F. (1997). Because every pupil counts: the success of the pupil monitoring system in The Netherlands. *Education and Information Technologies*, 2(4), 287–306.

Voogt, J., & Roblin, N. P. (2012). A comparative analysis of international frameworks for 21st century competences: Implications for national curriculum policies. *Journal of Curriculum Studies*, 44(3), 299–321.

Walker, A. A., & Engelhard Jr, G. (2014). Game-based assessments: A promising way to create idiographic perspectives. *Measurement: Interdisciplinary Research & Perspectives*, 12(1–2), 57–61.

Walton, S. (2005). The eVIVA project: Using e-portfolios in the classroom. *Qualifications and Curriculum Authority* website. Retrieved from www.qca.org.uk/downloads/10359_eviva_bett_2005.pdf

Weng, M. M., Fakinlede, I., Lin, F., Shih, T. K. and Chang, M. (2011). A conceptual design of multi-agent based personalised quiz game. Paper presented at *The Eleventh IEEE International Conference on Advanced Technologies* (pp. 19-21). Athens, USA.

Whetzel, D. L., & McDaniel, M. A. (2009). Situational judgment tests: An overview of current research. *Human Resource Management Review*, 19(3), 188–202.

Wiliam, D. (2007). Content then process: Teacher learning communities in the service of formative assessment. In D. Reeves (Ed.), *Ahead of the curve: The power of assessment to transform teaching and learning* (pp. 182–204). Bloomington, Indiana: Solution Tree.

Wilson, M. (2009). Measuring progressions: Assessment structures underlying a learning progression. *Journal of Research in Science Teaching*, 46(6), 716–730.

Wilson, M., Bejar, I., Scalise, K., Templin, J., Wiliam, D., & Irribarra, D. T. (2012). Perspectives on methodological issues. In E. Care, P. Griffin, B. McGaw, (Eds.), *Assessment and Teaching of 21st Century Skills* (pp. 67–141). New York: Springer.

Wilson, M., & Scalise, K. (2015). Assessment of learning in digital networks. In E. Care, P. Griffin, B. McGaw, (Eds.), *Assessment and Teaching of 21st Century Skills* (pp. 57–81). New york: Springer.

Wilson, M., Scalise, K., & Gochyyev, P. (2018). Learning in digital networks as a modern approach to ICT Literacy. In E. Care, P. Griffin, B. McGaw, (Eds.), *Assessment and Teaching of 21st Century Skills* (181–210). New York: Springer

Wood, M. (2009). Human Computer Collaborative Assessment—Access by Computer (ABC)—University of Manchester. *HEFCE JISC*.

Woolf, B. P. (2010). *Building Intelligent Interactive Tutors: Student-centered strategies for revolutionizing e-learning*. Massachusetts: Morgan Kaufmann.

Wouters, P., & van Oostendorp, H. (2017). *Techniques to Improve the Effectiveness of Serious Games*. New York: Springer.

Zapata-Rivera, D. & Bauer, M. (2012) Exploring the Role of Games in Educational Assessment. In J. Clarke-Midura, M. Mayrath, and D. Robinson (Eds.), Technology-based assessments for 21st century skills: *Theoretical and practical implications from modern research* (pp. 147–169). Charlotte, North Carolina: Information Age Publishing.

Zhuang, X., MacCann, C., Wang, L., Liu, L., & Roberts, R. D. (2008). Development and validity evidence supporting a teamwork and collaboration assessment for high school students. *ETS Research Report Series*, 2008(2), i–51.

Zuiker, S. J. (2012). Educational virtual environments as a lens for understanding both precise repeatability and specific variation in learning ecologies. *British Journal of Educational Technology*, 43(6), 981–992.

A publication of the
Office of Education Research,
NIE/NTU, Singapore © 2019

An Institute of

NANYANG
TECHNOLOGICAL
UNIVERSITY
SINGAPORE