Assessing soft skills in youth through digital games

VIRTUAL PAPER PRESENTED TO ICERI 2019
11 to 13 November, 2019

Sarah Pouezevara
Senior Research Analyst
RTI International
spouez@rti.org

Co-authors: Sonya Powers (RTI), Greg Moore (Independent), Carmen Strigel (RTI), Kathy McKnight (RTI)
From 2016 to 2018, a group of researchers at RTI International set out to explore the question of how to measure the extent to which education systems are equipping youth with workplace skills. One product of this research was a game aimed at measuring problem-solving skills using the game-based assessment approach. This study describes the process of developing and testing the problem-solving game, including methods and results from pilot activities in two countries. There is a need for monitoring the skills gap for secondary leaving youth, but haven’t been instruments for this purpose until recently. Most cognitive and personality tests are developed in western contexts and are not easily transferred to other international contexts. Through this research, the authors hope to demonstrate the potential for using stealth assessment to better measure soft skills internationally as well as encourage additional research and development of valid measures of soft skills for students worldwide.

Why Game-based assessment? Because traditionally, soft skills, including like problem solving have been measured via self-report, which is prone to bias and may lack the consistency needed for validity. Additionally, self-report measures generally require reading prompts and understanding likert scales, which may not be highly developed skills in certain contexts. Much of the literature on workforce readiness argues...
that skills necessary for the workplace are best assessed in a context not unlike that of a workplace, but this type of observational assessment is difficult to do at scale.

A more authentic measure would ideally be obtained if individuals could demonstrate skills in a highly structured environment with unobtrusive data collection methods. A growing body of knowledge in this domain is known as “stealth assessment” of skills within games or simulations. The advantages are that data is collected while the action is taking place; also, computer data collection and scoring can be automated to make complex skill inferences and gather more data than humans are capable of.
Problem solving is a central concern to teachers and schools everywhere. Success in school, work, and life no longer just requires retaining knowledge and executing familiar processes, but to transfer previously acquired knowledge to novel situations.

Evidence-Centered Design (ECD) was used as the framework for designing tasks and anticipating how to make sense of the data gathered by the game. The ECD framework uses domain analysis, domain modeling and a conceptual assessment framework to deconstruct a given skill into measurable tasks. A first step in selecting one assessment domain, out of all of the possible “21st century skills”, was to clearly define the various domains and determine which one may be suitable for stealth assessment.

A broad literature review and consultation with ECD and 21st century skill experts were carried out to cluster, deconstruct, define, and differentiate the skills commonly referred to as important for workforce readiness or employability in the 21st century [9]. The result was a set of skills that were reasonably distinct and well-specified. These were prioritized by importance (if they are predictive of work-related outcomes or in demand by employers) and if they are teachable and measurable. Problem
solving was selected for the first game to develop and was and defined as: The ability to activate prior knowledge (or to acquire new knowledge and apply it) to address a new problem through specific strategies. It consisted of cognitive, metacognitive, and motivational skill domains.
Because the context of a problem is likely to affect how an individual will succeed or fail at a problem-solving challenge, the design of the game needed to be based on a familiar situation. It also needed to be relatively universal, because the game modules were meant to be used in different countries, by children with very different academic and practical experiences.

For the process of co-development, RTI and the commercial game designer conducted a set of focus group sessions with a diverse group of youth (e.g., boys and girls, wealthy and poorer areas, in-school and out-of-school) around 15 years old in Morocco. Fifty youth participated in five development sessions during which participants were prompted individually to think of a scenario in their lives in which they had to overcome a challenge, and they were asked to describe the context, the problem, and the skills used to solve the problem. In small groups, they selected 2-3 of these real-life situations to develop into rich descriptions. By the end of the focus group sessions, 54 scenarios were proposed.

In the resulting game, called “Home Alone,” the player takes the role of an older sibling who needs to take care of two younger siblings and the house for a day while the parents are out. The player needs to observe
the siblings’ and animals’ needs, anticipate meal times, delegate tasks to keep the house clean and more. The tasks reflect the sub-skills from the framework established through the expert consultations, which are presented on the next two slides.
Players must make choices in advance about what types of meals to make, what ingredients to purchase at the store, and how to prioritize tasks. In identifying, finding, and obtaining meal ingredients or addressing animal needs, players also have a choice of delegating tasks to their younger siblings. The game also incorporates elements of other soft skills like dependability (task completion, time management), learning skills (accepting instruction); and interpersonal skills (self-regulation).
The game can be played offline as an application on the Android operating system or online using an Internet browser. As the player completes tasks in the games, variables are collected in the background and stored in a database. Game administrators can access this data in CSV file format from a web-based central dashboard. Data are logged in the device and then uploaded to the server as soon as it has an internet connection. Game play takes approximately 35 minutes.

<table>
<thead>
<tr>
<th>Skill domain: Facet measured</th>
<th>Operational Definition</th>
<th>Associated game behaviors</th>
</tr>
</thead>
<tbody>
<tr>
<td>(Problem Solving) Problem Identification</td>
<td>The ability to detect problems, and to understand and articulate the nature of the problem</td>
<td>Notice hungry sibling; Identify meal ingredients; Maintain livestock food.</td>
</tr>
<tr>
<td>(Problem Solving) Solution Identification</td>
<td>The ability to select appropriate techniques for addressing a problem</td>
<td>Character feed total; Get lunch ingredients; Get dinner ingredients</td>
</tr>
<tr>
<td>(Intrapersonal skills) Self-regulation</td>
<td>The ability to plan, monitor, and evaluate a course of action against a standard, as well as to control impulses that digress away from the course of action</td>
<td>Chasing chickens; Delegate tasks; Refuse Distractions</td>
</tr>
</tbody>
</table>
METHODS

- **Game-based assessment**: Skills are assessed through context-specific simulations.
- **Evidence-centered design**: Systematic process of choosing the constructs to assess, identifying the behaviors/performances that reveal them, how they will be captured, and how to score them (Mislevy, 2013)
- **Stealth assessment**: Data on behaviors/choice captured during simulation (Shute & Ke, 2012)
- **Offline and mobile first technology**: Assessments are run & data collected on mobile tablets that only require connectivity for sending data files, but not for playing.
- **Concurrent validity through triangulation**: Game play was analyzed along with other personality tests and self-assessments of each player
After the co-design process described earlier, RTI worked closely with the game designer to review the game as it was being created and to test functionality iteratively prior to initial user testing. In addition to standard software user interface testing processes, two pilots were organized to test game functionality in a live setting, to identify elements that needed to be fixed or revised, to gather data with which to test scoring approaches, and to establish concurrent validity.

The first pilot took place in Morocco. During sessions consisting of up to 10 individuals, 100 participants played the game, took a math and reading assessment, and responded to a 30-item self-report questionnaire comprised of five sub-scales, each with six items. The sub-scales were: academic mindsets, executive functioning, task completion, punctuality, and perseverance. Parents also completed the same 30-item questionnaire rating children’s problem-solving skills.

Both questionnaires were self-administered. The reading and math tests were primarily designed to determine the extent to which performance in the game may be dependent upon reading and math ability, both of which are required to some extent in order to play.
The testing in the Philippines was facilitated by a local research partner who recruited the students and obtained consent, managed childrens’ game play, and administered additional instruments. The sample consisted of fifty children from a special public science high school in Metro Manila. Unlike Morocco, the game was played on an internet-connected web browser in the school’s computer lab. The children did not complete reading and math assessments, but instead English proficiency, as determined by teachers, was a prerequisite for participation. Instead of parent surveys, teachers were also given a 5-item Likert survey of children’s skills. Children were also asked to self-assess, on a five-point scale the extent to which they understood the self-assessment questionnaire (a rapid measure of English proficiency), and to answer two questions about frequency of digital game use and their own assessment of how good they are at playing video games. For this pilot test, the researchers implemented the Raven’s Standard Progressive Matrices (Clinical Edition). A reduced set of 27 items was selected in order to reduce the overall time of administration to approximately the same amount of time as the game. Finally, youth completed a self-assessment questionnaire consisting of twenty-nine items consisting of six item sets related to industriousness, procrastination refrainment, control, caution, task planning, and perseverance. This measure of conscientiousness was administered because there appears to be some overlap in the subskills involved in problem-solving and the subskills involved in conscientiousness. Additionally, this measure provides a way to verify that the game is measuring what was intended by providing evidence of discriminant validity. The researchers expected the game-based measures of problem-solving to correlate more highly with Raven’s measure of problem-solving as opposed to the Likert self-report measure of conscientiousness.
The first pilot experience in Morocco indicated that players’ experience was largely positive, though several technical issues (e.g. game freezes, or crashing) hampered both user experience and data collection. Participants appreciated the game for its “realism” (face validity). Some players felt the game was difficult, in that it expected players to accomplish too many tasks by the time parents arrived home.

The language of the instructions in the game also appeared to be an issue for a number of participants. Many participants performed poorly on the reading and mathematics tests and data collected from the games (as well as anecdotal observation) suggested that most participants did not read the game instructions.

Results from the youth and parental questionnaires about their self-assessment of problem-solving skills tended to be related to themselves, but not with each other. This suggests that participants and parents either differ in their assessments of participants’ skills or differ in their understanding of the questions.

The game data was collected from the backend intact. Variables
performed as designed (i.e., they captured the specified information from the game modules). The variables collected by the game showed variation within the pilot population. While the variation is not perfect and many items were difficult for players (resulting in low scores on those items), there were no uniformly high or low scores across items. Finally, the researchers calculated alpha coefficients to determine the internal reliability of the scoring of skills. Several alpha coefficients were found to be acceptable (task completion, identifying solutions) and are close to 0.8, indicating a fair amount of internal reliability between the variables collected and included in the measurement of those skills. Other values, however, were quite low, and suggest that the variables included in the measurement of these skills might not map onto a single construct (but might measure different skills).
In the Philippines, despite prior trials, many of the children experienced a game freeze when attempting certain combinations of game play, which forced them to restart. The 50 players generated a total of 151 data records, but the number who made it far enough in the game to utilize the data in a meaningful way declined steeply midway through the game, with less than half of the group completing more than half of the game.

As in Morocco, the Philippines pilot data—even using partially completed sessions—found an acceptable amount of variation among the responses. An analytical process of aggregating variables, and checking for internal consistency identified a scoring framework that primarily used the number dirty spots in the house that the player cleaned by the end of the game. This variable is not dependent upon game completion so analysts used this indicator to compare against children’s self-assessment of conscientiousness and their problem-solving performance on Raven’s test.

Correlation of the Raven’s test with video game skill was low/moderate (0.23), indicating those who are more skilled at video games may be better problem solvers. The self-assessment total score, when correlated with other background variables found that the correlation with video
game usage and video game skill were low/moderate (0.25), the correlation with English proficiency was moderate (0.32), and the correlation with the Raven’s was low (0.19).

When the game’s outcome measure “number of dirty spots cleaned” was correlated with the validity measures and background variables, the pattern provided some evidence that the game based measure of problem-solving is more highly related to an external measure of problem solving than an external measure of conscientiousness, which was expected.

These results also confirm our expectation that a small positive relationship would be found between the game-based measure of problem-solving and the external measure of conscientiousness as the sub-skills have some conceptual overlap.

To further explore the relationship between the game indicator (percent of dirty spots cleaned), game skill, and problem-solving skill as measured by Raven’s, a multiple linear regression model was fit with Raven’s and video game skill as predictors of the percent of dirty spots cleaned. Neither video game skill nor student score on Raven’s were significant predictors of the game variable. Additionally, the total variance explained was small ($R^2 = 0.084$).
In both countries, the games performed largely as designed (i.e., children making decisions to complete tasks in the game); however, there were some persistent functionality issues that caused game freezes. Unfortunately, custom games can be unpredictable on different platforms. On the one hand, designing digital games to be cross-platform compatible introduces complexities; on the other hand, given how rapidly technology changes and the need to expand the use of tools as return on investment, it remains important to test across a wide range of platforms and devices even as the technology changes. This is a resource/cost-intensive process that researchers and game developers have to manage.

The ECD framework was successfully used to deconstruct the construct of ‘problem solving’ into measurable sub-skills, and the user-centered design with focus groups of target youth was an effective way to design a game that proved contextualized and relevant to the population in which the game would be deployed.

A challenge to measuring a skill that is in part defined by the ability to surpass difficulties using available resources is to fine tune the level of difficulty in the game such that it provides an opportunity to measure
differences, but is not a barrier to completing all tasks within the game.

In both Morocco and the Philippines, there were decreasing returns of efficiency with group testing. While larger groups afford the opportunity to assess more students in less time, they were also increasingly difficult to manage.

Another challenge to group-based testing of a digital game is that children, particularly in social-media rich environments like the Philippines, have a natural proclivity to work together and share progress, discuss tactics, and help each other along. This necessarily challenges the validity of a game designed to measure individual personality traits. The solution—placing children in individual desks with no interaction—creates an interesting dilemma where in order to diagnose and improve some 21st century skills, it may be required to place students in a traditional, individual seat work arrangement.

Finally, each of the external measures had its own limitations: In the Philippines, the teacher ratings were of little value because the teachers tended to rate the children all very highly. It is possible that the selection method of requiring strong English language ability had the effect of selecting highly functioning children into the sample, in which case teachers’ ratings were in fact accurate. But it is equally likely that in the Philippine context, it is culturally improbable that teachers would rate their children anything but positively. Also, given large class sizes, short school shifts, and a tendency towards teacher-centered instruction, teachers may not know their children individually. Similarly, in Morocco, parent ratings did not systematically line up with students’ self-assessments; it appears as if parents do not know their children or do not know how their children perceive themselves, rather than that they systematically disagree (either in under- or over-estimating) with their children’s self-perception.

This is the paradox inherent in the measurement of soft skills, is that if there were a perfect instrument to measure problem-solving skills then there would be no need to come up with a new one. The researchers did not seek high levels of correlation given that the game-based assessment was meant to improve on current approaches. However, the current body of knowledge does not provide evidence of what might be a desirable level of correlation. Therefore, finding an external measure with which to evaluate the validity of the game-based assessment is challenging. The researchers used instruments that were based on their knowledge of the best evidence available, but also modified them to meet the needs of the testing situation. In the end, the researchers did not yet have sufficient evidence to say that the game is a valid
measure of problem-solving skills.
One of the reasons why game-based assessment is an attractive alternative for international settings is that it is not highly dependent on the kind of literacy skills needed for certain self-assessment rating scales, it can be transferred easily from one context to another, and requires very little training and orientation to administer. In the Philippines, although our sample did come from a mostly urban environment of teens from select science-focused high schools, there were no instructions given for using the tablets or the game. The pilot test administration team only had to pass out the tablets and tell the children which game icon to open; from there they were able to play with little to no intervention from the team, except in the case of re-starting. Similarly, there is no need to train test administrators (except on basic logistics and standard, ethical research protocols) or encode survey data afterward. Thus the value of the methodology for scaling in a range of contexts is clear. However, establishing reliability and validity is extremely resource intensive.
<table>
<thead>
<tr>
<th>Name</th>
<th>Title</th>
<th>Organization</th>
<th>Email</th>
</tr>
</thead>
<tbody>
<tr>
<td>Carmen Strigel</td>
<td>Director, Technology for Education and Training</td>
<td>RTI International</td>
<td><a href="mailto:cstrigel@rti.org">cstrigel@rti.org</a></td>
</tr>
<tr>
<td>Sarah Pouvzevara</td>
<td>Senior Research Analyst</td>
<td>RTI International</td>
<td><a href="mailto:spouez@rti.org">spouez@rti.org</a></td>
</tr>
</tbody>
</table>

![CurrantMobile™ logo]