Game-Based Assessment of Persistence

Kristen E. DiCerbo

ABSTRACT
Interest in 21st century skills has brought concomitant interest in ways to teach and measure them. Games hold promise in these areas, but much of their potential has yet to be proven, and there are few examples of how to use the rich data from games to make inferences about players’ knowledge, skills, and attributes. This article builds an evidence model for the assessment of persistence from Poptropica, a popular commercial game for children. Task persistence is an important skill related to successful school and work outcomes, particularly given new, complex tasks requiring sustained application of effort. Evidence extracted from log files of the game was used to identify players with a particular goal and then create a measure of persistence toward that goal. The results support the ability to create an assessment argument for a game-based measure of persistence.

Keywords
Game-based assessment, Persistence, Measurement models, Educational data mining

Introduction

The digital revolution is fundamentally changing the nature of work, the problems workers are asked to solve, and therefore the types of skills needed in today’s world. Today’s workers must be able to apply their knowledge to more complex tasks, generate creative solutions to multi-faceted problems, and navigate interwoven systems. Employers indicate that skills such as problem solving, communication, collaboration, and creativity, as well as personal attributes such as adaptability, persistence, and resilience are becoming more important, and have been labeled “21st century skills” (Casner-Lotto & Barrington, 2006; Fadel, 2011). However, employers also indicate that employees often lack these essential skills (American Management Association & Partnership for 21st Century Skills, 2010), leading to a push for them to be taught and assessed in schools.

However, these skills, by their very nature, do not lend themselves well to traditional methods of assessment. Most traditional tests present highly decontextualized individual items to learners. The 21st century skills and attributes of interest require application in context as part of complex tasks for accurate measurement. In addition, traditional assessment often interrupts the learning process in order to gather information and does little to motivate the learner to put forth effort, further jeopardizing our ability to gain valid estimates of skills and attributes. In response to these concerns, there has been growing interest in, and investigation of, the use of games to assess 21st century skills (Shaffer et al., 2009; Shute, 2011).

Games are attractive as assessment tools for a number of reasons. First, they allow us to make observations in contexts closer to those in the real world, creating the complex scenarios required to evaluate the application of knowledge and skills. Second, we know that games are engaging and motivating and that assessments are more valid when students are more motivated by them (Schmit & Ryan, 1992; Sundre & Wise, 2003). Third, we know that the vast majority of students (for example, 97% of teens aged 12-17 in the United States) already play digital games (Lenhart et al., 2008). This means we do not have to stop students’ daily activity in order to gather assessment information. Rather, we can tap into the digital ocean of data already being produced by their online activity to identify and accumulate evidence about what players know and can do from that (DiCerbo & Behrens, 2012). Finally, games and assessment share a similar process loop of activity presentation, activity completion, evidence identification, evidence accumulation, and presentation of the next activity (Behrens, Frezzo, Mislevy, Kroopnick, & Wise, 2006), making games ready targets for assessment efforts.

However, before the potential of games as assessments of 21st century skills can be realized, there are a number of practical challenges to be addressed. New interactive digital experiences such as games elevate the importance of micro-patterns in data which often reflect variation in strategy or evolving psychological states. Highly granular data about minute human-computer interactions are now available in vast quantities. While the richness of the data holds promise, standard methods by which to turn these data into inferences about knowledge, skills, and attributes are not well-developed. We must be able to identify evidence of our constructs of interest in large files that log the small details of in-game events and we need models for the accumulation of this evidence into estimates of students’ skill
proficiency. Many of our current models of assessment design and delivery will have to be modified, and some completely discarded and re-imagined, in order to conduct game-based assessments of 21st century skills.

Evidence-centered design

A framework to assist with both the conceptualization and implementation of game-based assessment can be found in Evidence-Centered Design (ECD; Mislevy, Steinberg, & Almond, 2003). ECD assists us in developing the assessment argument (Whitely, 1983). It specifies an evidence model that links the output of a learner/player, in this case the data stream from their interaction with a digital environment, and the skill or attributes we are interested in. The evidence model consists of evidence rules and measurement models. Evidence rules refer to the means by which we select particular elements from a student’s work product and apply scoring rules to obtain observable values. Measurement models provide us with the method of combining these observable values to make inferences about our constructs of interest (Mislevy et al., 2003).

In practice, evidence rules and measurement models are often more complex for digital learning environments like games than for traditional assessments. In a digital learning environment, it is not always clear which aspects of the submitted work products (e.g., causal maps, reports of experiments, configured computer networks) should be attended to or how they should be weighted and aggregated to form an overall estimate of skills, knowledge, or attributes.

Persistence

This paper describes efforts to define an evidence model for the attribute of persistence. Task persistence is defined as continuing with a task despite obstacles or difficulty. In the cognitive literature, persistence is generally classified as an element of executive function and thought to be related to self-regulated attention, cognition, and behavior (Anderson, 2002). Persistence may not seem like a distinctly 21st century skill, given that there was a historical review of the literature on measurement of persistence written in 1939 (Ryans, 1939). However, it is often enumerated in lists and discussions of 21st century skills and attributes (Fadel, 2011; Pellegrino & Hilton, 2012), because jobs in the 21st century are increasingly complex, requiring sustained application of effort to complete multifaceted tasks (Andersson & Bergman, 2011).

Persistence is of particular interest and importance because persistence at a young age has been shown to be predictive of many academic and employment outcomes, including adult educational attainment, income, and occupational level (Andersson & Bergman, 2011). The relationship between persistence and academic achievement has been repeatedly documented (Boe, May, & Boruch, 2002; Deater-Deckard, Petrill, Thompson, & DeThorne, 2005; McClelland, Acock, Piccinin, Rhea, & Stallings, 2012).

It is not just persistence with academic tasks that predicts persistence with other academic tasks. Two-year-olds who spent more time trying to open a plexiglass box containing a toy were found to have fewer behavior problems and were more likely to complete school work at age 5 (Sigman et al., 1987). McClelland et al. (2012) report that parents’ ratings of their children’s persistence with difficult toys predicted college completion by age 25, suggesting valid, reliable measures of persistence may help us monitor and intervene with an aspect of learners that can significantly impact their future success.

In order to determine the indicators of persistence to look for in game play, we can examine how other researchers have operationalized it. When directly observing behavior, some have counted the number of times a task is attempted (Foll, Rascle, & Higgins, 2006) while others measure the amount of time during which a child exhibits task-directed behavior (Sigman et al., 1987). Importantly for this paper, Shute and Ventura (2013) defined four indicators to measure persistence in a digital game: time spent on a solved problem (by difficulty), time spent on an unsolved problem (by difficulty), number of restarts for solved problems (by difficulty), and number of re-starts for unsolved problems (by difficulty). It appears the commonalities across studies are time spent on difficult tasks, task completion and attempts after failure.
Purpose and research question

This paper describes research that attempted to determine whether a common factor could be identified under game actions thought to indicate persistence, as revealed by data captured in log files from Poptropica®, a popular commercial game for children. It describes the scoring model and the evaluation of a measurement model that result in a measure of persistence.

Method

The game

Poptropica® is a virtual world in which players explore “islands” with various themes and overarching quests that players can choose to fulfill. Players choose which islands to visit and navigate the world by walking, jumping and flipping. The quests generally involve 25 or more steps, such as collecting and using items, usually completed in a particular order. For example, in Vampire’s Curse Island (see Figure 1), the player must rescue her friend Katya who has been kidnapped by Count Bram. In order to do this, she must navigate to the Count’s castle (eluding wolves on the way), discover the potion for defeating the vampire, identify and mix the potion ingredients, hit the vampire with a potion-tipped arrow from a crossbow, and then find Katya. Apart from the quests, players can talk to other players in highly scripted chats, play arcade-style games head-to-head and spend time creating and modifying their avatar.

![Figure 1. Screenshot from Poptropica’s vampire’s curse island](image1)

![Figure 2. Example of log file](image2)
The backend of the game captures time-stamped event data for each player. Events include, for example, the completion of quest-related steps, modifying avatars, collecting objects, and entering new locations (see Figure 2). On an average day 350,000 players generate 80 million event lines. Players can play the game anonymously and for free or, with purchase of a membership, and gain access to premium features such as early access to new quests.

Persistence in general requires the presence of a difficult task (Shute & Ventura, 2013), given that easy tasks will not provide the challenge and failure that will allow for the exhibition of persistence. One way to measure difficulty is the percent of players who successfully complete a task. Across islands on Poptropica, completion rates range from 3% to 9%, indicating that these are not easy tasks to complete.

Participants

The final sample consisted of 892 players, 51.2% females and 48.8% males, which is consistent with the near-even gender divide in the overall Poptropica player population. Players ranged in age from 6 years old to 14 years old, with an approximately normal distribution across ages, peaking at age 10 (18.6% of the sample).

The final sample was selected from a larger initial sample. First, a group of players who had at least 500 game events in one of two days was selected. Given that persistence is defined in terms of work towards a particular goal (Meier & Albrecht, 2003), we quickly determined that in order to investigate persistence towards a goal we had to first identify people who were working toward that goal; it would be inaccurate to describe someone as lacking persistence when in fact they were being persistent toward another goal. In this research, we defined persistence as persistence toward completing quests in the game. However, players of the game have varying goals, including social interaction and creating avatars. Therefore, a sample of players with the goal of completing a quest was created by selecting players who completed at least the first purposive action required to complete the quest.

A final examination of distributions by island resulted in the removal of three islands from the dataset due to their low numbers of events required to complete a quest, resulting in outlying patterns of players on quest event-related indicators. Finally, all individuals who had three islands that they made “serious” attempts at completing were combined to form the final sample.

Selection of indicators

The creation of indicators from the log file data was based on the salient elements of measures of persistence in previous research, summarized above, and the data available from Poptropica. Previous research indicated that measurement of persistence required tasks that would be difficult for those attempting them, in order to observe responses to this difficulty. In Poptropica, fewer than 10% of those beginning quests go on to complete them and the number of YouTube walkthroughs and hint pages published by players suggest that the quests are challenging to solve. In order to measure the response to this difficulty, a number of potential indicators were considered in line with the time spent, tasks completed, and attempts after failure. Specifically four potential indicators from each quest were investigated: time spent on quest events, number of quest events completed, maximum time spent on an individual quest event successfully completed, and time spent on the last event prior to quitting the island (unsuccessful quest event). Data was not captured in the game in a way that would allow for a count of number of attempts after failure; this could only be inferred from the time from the completion of one event to the completion of the next event. The variation and relationships among these four indicators were investigated with a small pilot sample. Unfortunately, the data from individual events, both successful and unsuccessful, proved to be highly variable and unreliable as indicators. Therefore, we settled on two indicators per quest: the total time spent on quest-related events and the number of quest events completed. These indicators were computed from the log files for each of the three quests each player attempted to “seriously” complete.
Results

Data cleaning

Examination of the distributions of the indicator variables indicated that they were highly skewed. Given the normality assumptions of the procedures in the study, various transformations of the variables were explored and it was determined that a logarithmic transformation of the number of quest events and a square root transformation of the total time variable resulted in more Gaussian distributions (see Figure 3). The variables were then standardized by island into z-scores in order to account for between-island differences in number of events and time required to complete them.

![Figure 3. Logarithmic transformation of Quest Events indicator](image)

Confirmatory factor analytic model

A confirmatory factor analytic model was run in order to evaluate whether a common factor explained variance across quest indicators. In confirmatory factor analysis (CFA), we specify a model of how the variables are related a priori, and then examine whether the data match what the model would predict. In the case of this research, if these measured variables have a pattern of variance and covariance that is in accordance with what our specified model would predict, that would indicate there is a common construct underlying this group of indicators, and we can then discuss what that construct is and how we can create scores for it.

The first step in a confirmatory factor analysis is the examination of the pattern of correlations. If our indicators all measure a single underlying construct, we would expect moderate to large correlations between them. Table 1 presents the correlations between the indicators for each of the three islands. Correlations are estimates of linear relationships. The raw, non-normal variables do not have a linear relationship. Therefore, the correlations of the transformed variables provide a better estimate of the correlations between the underlying constructs than the raw data.

<table>
<thead>
<tr>
<th></th>
<th>I1 Quest Events (Log)</th>
<th>I1 Total Time (SQRT)</th>
<th>I2 Quest Events (Log)</th>
<th>I2 Total Time (SQRT)</th>
<th>I3 Quest Events (Log)</th>
<th>I3 Total Time (SQRT)</th>
</tr>
</thead>
<tbody>
<tr>
<td>I1 Quest Events (Log)</td>
<td>1.00</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>I1 Total Time (SQRT)</td>
<td>.64</td>
<td>1.00</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>I2 Quest Events (Log)</td>
<td>.48</td>
<td>.47</td>
<td>1.00</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>I2 Total Time (SQRT)</td>
<td>.49</td>
<td>.64</td>
<td>.64</td>
<td>1.00</td>
<td></td>
<td></td>
</tr>
<tr>
<td>I3 Quest Events (Log)</td>
<td>.42</td>
<td>.39</td>
<td>.54</td>
<td>.55</td>
<td>1.00</td>
<td></td>
</tr>
<tr>
<td>I3 Total Time (SQRT)</td>
<td>.38</td>
<td>.48</td>
<td>.56</td>
<td>.69</td>
<td>.80</td>
<td>1.00</td>
</tr>
<tr>
<td>Median</td>
<td>15</td>
<td>688</td>
<td>11</td>
<td>397</td>
<td>8</td>
<td>215</td>
</tr>
<tr>
<td>IQR</td>
<td>12</td>
<td>759</td>
<td>10</td>
<td>505</td>
<td>7</td>
<td>305</td>
</tr>
</tbody>
</table>

*Note. I1=Island 1, I2=Island2, I3=Island 3, SQRT = Square Root, IQR = Inter-quartile range*
A confirmatory factor analytic model was specified with the two indicators of persistence from each of the three islands loading onto a single factor. The error terms for the pairs of indicators from the same island were each correlated because these measures are not independent (completing more events requires more time). Given this hypothesized model, we now want to evaluate it to determine whether it is likely a true representation. In order to do this, we compare the variance-covariance patterns the model would predict to the variance-covariance patterns in the actual data. To make this comparison, we use fit indices that estimate how well the data fits the model. However, because these indices can be sensitive to things like sample size and the size of the model, there are a number of different indices that need to be examined to gain a complete picture of adequacy of fit.

The model fit statistics for this model are displayed in Table 2. The chi-square is significant, which is not desirable, but chi-square is notoriously sensitive to sample size (Kline, 2005), which led to the consideration of alternative fit indices. One way to use the chi-square is to divide it by the degrees of freedom, with values less than 3 indicating good fit (Schermelleh-Engel, Moosbrugger, & Müller, 2003). The Comparative Fit Index (CFI) is an estimate of the improvement of the model over a null model, with a correction for sample size. The Tucker-Lewis Index (TLI; also called the Non-Normed Fit Index) is also an incremental fit index, but with a correction for model complexity. In both cases, values over .95 are indicative of good fit (Schermelleh-Engel et al., 2003; Sivo, Fan, Witta, & Willse, 2006). The Root Mean Square Error of Approximation (RMSEA) is an absolute measure of approximate fit in the population and cutoffs of .01, .05, and .08 have been suggested for excellent, good, and mediocre fit, respectively (MacCallum, Browne, & Sugawara, 1996). The Standardized Root Mean Square Residual (SRMR) is a standardized summary of the average differences between the observed and model-implied covariances. In this case, values less than .10 are indicative of good fit (Schermelleh-Engel et al., 2003).

Examination of the fit of the original model indicated that only the CFI and SRMR met criteria for good fit. Examination of the modification indices suggested that the three indicators of the number of quest events should be correlated. This appeared to be a reasonable modification, given that these three measures could be considered the same measure given across three time periods. Reddy (1992) cautions that ignoring correlated measurement error may lead to inaccurate estimates of coefficients. Correlating these errors resulted in a model with the fit in the second row of Table 2, which indicates good to excellent fit across all indices. This model is displayed in Figure 4.

All of the unstandardized coefficients in the model are significant. The standardized coefficients for the error terms for each indicator provide the percent of variance for each indicator unexplained by the model. For most of the indicators, the persistence variable by itself still leaves a portion of variance unexplained; however, the six indicators do share a significant portion of variance. Examination of the loadings indicates that the time variables appear to have slightly stronger loadings than the completion variables. In addition, the correlation between the two indicators for each island is moderate and significant, indicating that the time taken and events completed on a given island are related above and beyond the variance in persistence. This correlation might indicate the effects of the context of the island, such as whether the island theme was of interest to the player.

![Figure 4. Standardized solution for measurement model of persistence](image-url)
### Table 2. Model fit summary

<table>
<thead>
<tr>
<th>Model Type</th>
<th>$\chi^2$</th>
<th>df</th>
<th>$\chi^2$/df</th>
<th>CFI</th>
<th>TLI</th>
<th>RMSEA</th>
<th>SRMR</th>
</tr>
</thead>
<tbody>
<tr>
<td>Original Model</td>
<td>88.09</td>
<td>6</td>
<td>14.68</td>
<td>.97</td>
<td>.93</td>
<td>.13</td>
<td>.03</td>
</tr>
<tr>
<td>Modified Model</td>
<td>5.16</td>
<td>3</td>
<td>1.72</td>
<td>.99</td>
<td>.99</td>
<td>.03</td>
<td>.01</td>
</tr>
<tr>
<td>Cross-Validation – configural invariance</td>
<td>7.75</td>
<td>5</td>
<td>1.55</td>
<td>.99</td>
<td>.99</td>
<td>.02</td>
<td>.01</td>
</tr>
<tr>
<td>Cross-Validation – metric invariance</td>
<td>15.86</td>
<td>11</td>
<td>1.44</td>
<td>.99</td>
<td>.99</td>
<td>.03</td>
<td>.02</td>
</tr>
</tbody>
</table>

Note. CFI = Bentler Comparative Fit Index; TLI = Tucker Lewis Index; SRMR = Squared Root Mean Square Residual

Although the fit of the re-specified model is very good, the post-hoc modifications run the risk of capitalizing on chance and sampling error in the initial sample. For this reason, an independent sample of 240 players was randomly selected from a different day of play to cross-validate the results. A factorial invariance approach was used, first assessing configural invariance, or the notion that there is the same number of factors (one in this case) and the same free and fixed parameters across both groups. As the fit of this model, displayed in Table 2, indicates, configural invariance was supported, $\chi^2(5) = 7.75, p=.26$ suggesting that the error parameters correlated in the model modification step also result in good fit in the independent sample. Next, metric invariance was tested by fixing the factor loadings between the two samples to be equal. Again, invariance was supported, $\chi^2(11) = 15.86, p = .15$. The change from the configural model to the metric model, $\Delta\chi^2(5) = 8.12, p = .15$, supports the equivalence of the factor loadings across the two samples.

The model can also be used to create factor score estimates for each individual. A number of different methods exist for this computation. In this case, Thurstone’s regression method was employed, in which the observed standardized values are multiplied by regression coefficients (which are obtained by multiplying the inverse of the observed correlation matrix by the matrix of factor loadings. This results in a standardized metric, much like a z-score, with a mean around 0, but where the standard deviation depends on the application. In this case, the overall factor scores have a mean of 0 and a standard deviation of .36.

**Relationship to grade**

These individual factor scores can be combined to examine group trends as well. Although played anonymously, Poptropica players enter their grade prior to playing the game. Persistence has been shown to increase over the ages spanned by these grades (Hartshorne, May, & Maller, 1929; Lufi & Cohen, 1987), likely as a typical developmental change. Therefore, we computed factor scores for all players in the original sample and examined their relationship to age. As shown in Figure 5, there is a general upward trend in scores across grades, rising from first grade ($M = -.13, SD = .37$) to ninth grade ($M = .13, SD = .34$), for a change of .72 deviations across the nine grades.

![Boxplot of persistence factor scores by grade](image)
Reliability

The internal consistency reliability of the scale, as measured by alpha, was .87. There are essentially only six indicators in this measure so a clear way to increase the estimate here would be to add observations from more islands. However, this would also reduce the number of players for whom we could create estimates because fewer players make serious attempts at four or more islands. The reliability when leaving each indicator out in turn is presented in Table 3.

<table>
<thead>
<tr>
<th></th>
<th>Alpha when removed</th>
</tr>
</thead>
<tbody>
<tr>
<td>I1 Quest Events (Log)</td>
<td>.85</td>
</tr>
<tr>
<td>I1 Total Time (SQRT)</td>
<td>.86</td>
</tr>
<tr>
<td>I2 Quest Events (Log)</td>
<td>.85</td>
</tr>
<tr>
<td>I2 Total Time (SQRT)</td>
<td>.83</td>
</tr>
<tr>
<td>I3 Quest Events (Log)</td>
<td>.85</td>
</tr>
<tr>
<td>I3 Total Time (SQRT)</td>
<td>.84</td>
</tr>
</tbody>
</table>

*Note. I1=Island 1, I2=Island 2, I3=Island 3, SQRT = Square Root*

Discussion

The purpose of this research was to demonstrate the use of data from game play log files to assess a 21st century attribute. Specifically, persistence, or continuing in the face of difficulty, was assessed with a combination of completion and time indicators. Examination of the existing literature on persistence indicated that time spent on difficult tasks, task completion, and continuing after failure were the most common measures of persistence. Similar indicators were created from the log files of the Poptropica game and tested via confirmatory factor analysis. The results suggest there is likely an underlying factor that explains variance across these indicators, and cross-validation on an independent sample confirms the structure and factor loadings. Finally, scores created from these indicators increase across grade levels as we would expect, providing preliminary evidence of validity based on relationships to external variables.

The model fit indicated here reveals that the pattern of variances and covariances in the data very closely fit the model specified. This suggests it is reasonable to discuss a construct that underlies the observed measures. However, in confirmatory factor analysis, one must be cautious in interpretation due to the presence of equivalent models (Kline, 2005). In any confirmatory factor analysis, there are a number of models that are statistically equivalent, that is, will produce the same fit because the models themselves imply the same variance-covariance relationships. It is up to the researcher to provide justification for selecting among these models. In this research, an equivalent model would be to specify a hierarchical model in which the two indicators from each island for their own island-specific measures of persistence that then are combined into a higher-order persistence measure. This model would produce identical statistical results to the model here, and might be chosen if we were interested in a particular research question around comparing islands. However, since we are not, this hierarchical model simply adds more complexity without contributing explanation, so the more parsimonious model was chosen. As Kline notes, “if a single-factor model cannot be rejected, then there is little point in evaluating more complex ones” (2005, p. 211).

A key in the creation of measures is to determine whether they adequately represent the breadth of the construct of interest. Construct underrepresentation occurs when tasks in an assessment fail to include important dimensions of the construct. As described in the literature, persistence is a unidimensional construct and is often measured by a single indicator in research. In this case, we measured both time spent and completion of difficult tasks. As persistence is defined as continuing in the face of difficulty, this appears to cover the breadth of the construct. However, there is also still the clear need for further validation of the measure. Evidence from relationships to external sources, such as rating scales or measures in other settings, are needed to establish that results obtained via this game method correspond to results obtained by other means, providing further support that this is in fact measuring persistence. This lack of validation from external sources limits the conclusions that should be drawn from the measure at this time. However, there are few examples of combinations of indicators from log files of games to measure latent constructs, so this research can be taken as an example of the preliminary work that needs to be done to complete this task.
One important element to note in this analysis is the identification of a sample of players all working toward the goal of completing a quest. Efforts to model the data on the overall player population indicated a general lack of correlation among our indicators of interest. It was the realization that persistence is progress toward a particular goal and then the measurement of just those with a particular goal that resulted in a model with good fit to the data. In fact, the identification of a desired outcome or goal has been identified as a first step in a persistence process (Meier & Albrecht, 2003). This point is important for at least two reasons. First, when researching assessment in games, we must realize that many games are very open-ended spaces and players may be pursuing a variety of goals in game play, including, for example, achievement in the game, exploration, and socializing with others (Bartle, 1996). If any type of inference is to be drawn based on players’ behavior, it is important that their motivations and goals in the game be clear. Second, many people are interested in persistence toward larger goals outside games, particularly completing secondary and post-secondary schooling (e.g., Brown et al., 2008; Kuh, Crue, Shoup, Kinzie, & Gonyea, 2008). It is important to remember that not everyone enrolling in post-secondary education has the same goal (Provansnik & Planty, 2008). Any research on persistence should take the issue of goals into account.

Games for assessment, learning, and fun

This research was completed using a commercial game that was not designed for assessment. This is useful as an exemplar that demonstrates the potential of all kinds of games for measurement and the ability to use data captured from every day interactions to make inferences about players. However, it also led to limitations based on the types of data that were available. Given that the data to be collected from the game were defined long before this research began, events that might have improved this measurement were not tracked. For example, it would have been preferred to have a record of each attempt and failure individuals made at an event. When the player is trying to get past a wolf, it would have been nice to have a count of how many times the wolf chomped the player and how many times the player tried again to get past. This data was not available, only the eventual success at getting past the wolf is recorded along with the time spent trying to do so. This time can serve as a proxy for the number of tries (more tries is likely related to more time) but is not a direct measure. Time can also be subject to things like distraction, day dreaming, and mental processing speed.

This lack of identification of appropriate evidence for our constructs is one danger of not addressing key assessment constraints from the beginning of game design (Mislevy, Behrens, DiCerbo, Frezzo, & West, 2012). If we are interested in making inferences about players, a better approach is to bring together game designers and assessment experts from the beginning of the design process. Early design sketches should include not just game features, but preliminary assessment arguments. Essentially this partnering from the beginning allows for the building of better links between what players do, what elements of their work product are collected, how they are scored, and how we accumulate the information to make inferences.

Researchers also need to consider how learning and assessment interact in games. Discussions of gaming in education often focus on the evidence (or lack thereof) that games can enhance learning. This paper has not addressed that question. Rather, it has demonstrated the use of games to gather information about a student attribute that would be difficult to assess with traditional assessment. However, a next step for this research might be to investigate whether persistence changes over the course of continued game play.

Future research, apart from the needed external validation of the measure described here, might then focus on how games might be designed to foster and/or train persistence. It might be hypothesized that elements of game play reward players for persisting in their efforts, thereby increasing the likelihood that they will persist more in future interactions (Ventura, Shute, & Zhao, 2012). Although its definitions as a personality trait suggests for some that persistence might be immutable, research has shown that it is influenced by a variety of task and personal factors (Kamins & Dweck, 1999; Reiher & Dembo, 1984; Schunk, 1983; Thomas & Pashley, 1982). It appears that feedback that encourages students to attribute success to hard work as opposed to internal factors, has shown to be effective in increasing task persistence in other environments. Given games already are efficient at providing feedback to players, designing them to give this specific type of feedback would be an interesting challenge. Finally, the relationship between persistence and factors such as prior content knowledge and prior experience in a given learning environment influence persistence.
Games and learning analytics

While games have many unique qualities, they represent just one digital environment that can collect data. Today’s learning management systems allow us to collect data from a wide swath of sources. This can include usage information about logins to learning management systems, posts to discussion boards, interactions in simulations, along with traditional assessments. All of this ubiquitous data, collected in a largely unobtrusive manner, allows us to make inferences about students in new ways (DiCerbo & Behrens, 2012). The emerging field of learning analytics focuses on gathering, analyzing, and applying all of these new data types to the issue of understanding and improving learning (Siemens & Gasevic, 2012).

Learning analytics includes the mapping of knowledge domains and then the evaluation of learner activity in relation to those maps. Assessment becomes a continual activity, as new data is collected, it is used to update a students’ profile of knowledge, skills, and abilities. Content and intervention, when needed, can be recommended and delivered in response to this changing profile, rather than in lock-step as it is currently (Siemens & Long, 2011). Games then, may provide just one piece of evidence that can provide information into such a profile. In the case presented here, a measure of persistence from a game like Poptropica might be combined with a persistence measure from viewing course videos and one from a homework system. A measurement of low persistent might trigger an alert to an instructor, and might also inform an instructional system to appropriately adjust feedback given to the student.

Concluding thoughts

The promise of game-based assessment is that we can use data from in-game actions to make inferences about players’ knowledge, skills, and attributes, allowing us to make use of information in the ocean of data produced by daily digital interactions with software (DiCerbo & Behrens, 2012). In order to do this, we need to be able to identify evidence from the log files produced by these interactions and accumulate that evidence to produce valid, reliable measures. This paper demonstrates one promising effort to create a model of persistence from young players’ actions in Poptropica using a combination of data mining and factor analytic techniques. While the loadings and indicators presented here are specific to Poptropica, the ideas could be applied across settings. Indicators of time spent on and progress through difficult tasks can be gathered from a variety of games and other digital activities. These indicators can be standardized and combined through factor scores or other methods of aggregation. To summarize, for others interested in assessing persistence, this work suggests, they should (1) clearly identify the goal individuals are working toward, (2) measure how much progress was made toward the goal, and (3) measure how long players spent trying to complete difficult tasks. These measures or indicators can be combined with relatively common statistical techniques, including factor analysis. The continued exploration and combination of new and existing methods to extract information from data streams may fundamentally change how we think about assessment.

Acknowledgments

The author wishes to thank Dan Murphy, David Shin, Steve Ferrara, and the anonymous reviewers for their suggestions which greatly improved this work.

References


