Lessons Learned from Testing a Children's Educational Game through Web Deployment

Michael G. Christel, Scott M. Stevens, Bryan S. Maher, Matthew Champer, Samantha Collier, Ricardo Merchan Benavides, Sean Brice
Entertainment Technology Center, Carnegie Mellon University
Pittsburgh, PA USA
christel@cs.cmu.edu

ABSTRACT
Beanstalk is an educational game that teaches balance-fulcrum principles, targeting children ages 5-11. Four versions of the game were deployed on the web through the educational portal Learning.com. Two of the versions incorporate non-player characters that offer opportunities for socio-emotional learning (SEL). Two of the versions include a scientific process of "predict-observe-hypothesize-explain" (POHE), which in effect is in-game testing. This paper reports on a first look into the game logs collected from child players via the portal, allowing for a 2x2 SEL crossed with POHE analysis. Lessons learned from testing variants of an educational game are shared, as well as conclusions drawn from the logs. Surprisingly, the version with both SEL and POHE caused children to stop playing the game sooner. The POHE performance was weakest for the players in the SEL treatment. Rather than help with the science educational goals, SEL may have diminished the experience.

Categories and Subject Descriptors
H.5.2 User Interfaces: Evaluation/methodology; K.3.1 Computer Uses in Education: Computer-assisted instruction (CAI); K.8.0 Personal Computing: Games.

General Terms
Experimentation.

Keywords
Educational game; analytics; early childhood science education; game metrics; adaptive learning; game design.

1. INTRODUCTION
Beanstalk is a Unity web player game developed to teach children ages 5-11 scientific principles, socio-emotional learning (SEL), and the scientific process of hypothesis testing through experimentation and observation. The game integrates cartoon-style artwork and animations, sound effects, and voice-over dialog into an experience that can be won to its final stages in 30-45 minutes. The game is based on Siegler's cognitive development work with a balance scale [8, 9], teaching principles governing the sum of cross products rule that can be used to determine whether a scale will balance, given a particular configuration of weights on each side of the fulcrum. The game levels are designed to help children progress through four increasingly sophisticated mental models, i.e., "Rules" [3], identified by Siegler: (1) paying attention to weight, not distance (i.e., side with more weight goes down); (2) considering distance, but only when weight is equal on both sides; (3) considering both weight and distance, with cues agreeing with each another; (4) considering both the amount of weight and distance of weights from the fulcrum; if the cues suggest different outcomes, the sum of cross products rule is applied. The level design respects the Lens of Flow as outlined in Schell's game design book [6] founded on earlier work by Csikszentmihalyi [2]. Level complexity increases to let the child player enjoy a rewarding experience to ideally remain engaged and feel a sense of achievement without undue frustration.

Beanstalk also optionally addresses SEL and giving the child player experience in constructing hypotheses and explanations, a scientific practice called out in the National Research Council report on K-12 science education [5]. SEL aspects of the game include seeking assistance when encountering a problem, cooperating to accomplish a joint task, and solving problems through interactions. Non-playable characters (NPCs) in *Beanstalk* allow for the practice of SEL and gaining skill in persisting through challenging levels, asking for help, cooperating, and discussing. Figure 1 shows an interface from an SEL version of *Beanstalk*, with Chicken and Crow "help' buttons to call up assistance. These NPCs also fly around in the game space with additional visuals and aural voice-over encouragement and commentary.

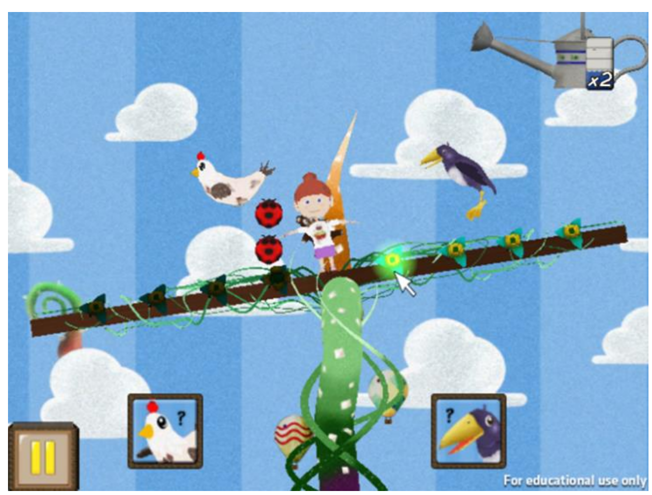

Figure 1. Screen shot of SEL version of *Beanstalk*, where chicken and crow chat with player. Water inventory of 2 supports the mirror solution of 2 added to highlighted pod.

SeriousGames'14, November 07 2014, Orlando, FL, USA

http://dx.doi.org/10.1145/2656719.2656726

During development of *Beanstalk*, there was concern over creating a game centered on the balance scale while also tackling SEL and dialog chains of "predict-observe-hypothesize-explain" (POHE). Would the added dimensions of SEL and POHE help or hurt with progression through Seigler's mental models and success with the problems? A formative study was conducted in a set of classrooms with 64 children ages 6-9 using a 2x2 SELxPOHE experimental design [1]. That work found that children using SEL treatments had significantly more losses and fewer wins in problem sets, with an average of 13.1 problems seen per player in a 20-minute game session. For SEL treatments, the child faced the unexpected challenge of a "cooperative" game level in which they had to ask the chicken to lay an egg on the beam, rather than water the spot directly, because that level forced cooperation by not providing enough water to solve the level without the chicken's help. Game logs showed that this forced SEL interaction did not work, needing a gentler introduction: players in the SEL treatments failed such levels repeatedly. Children in POHE treatments had significantly fewer losses (but no difference on wins), perhaps due to instruction provided during predict/hypothesize/explain cycles noting the relevance of distance and weight to the balance problem. It may also be due to more time being spent in POHE cycles and less on balance problem levels [1]. *Beanstalk* was revised to allow but not force cooperation, i.e., chicken egg-laying was optional but not required to pass levels in SEL treatments. The next round of testing was set up to occur "in the wild" through web deployment of *Beanstalk* to an educational portal: Learning.com. The game automatically loads up one of four configurations randomly when started by a new player: SEL and POHE, SEL with no POHE, no SEL and POHE, neither SEL nor POHE. This paper reports on the data collected for these four versions from logs produced from May 2013 to June 2014.

2. BEANSTALK: GAME FLOW

Lessons drawn from Schell's game design lenses [6], confirmed by playtests with children conducted early in *Beanstalk*'s development, emphasized the importance of narrative in educational games targeting 5-11 year olds. Without a framing story, the experience might be seen as a series of tests or exercises and lose the player's interest. With a good story, both girls and boys can be motivated to keep working through increasingly complex levels to arrive at an end goal. For *Beanstalk*, the story begins with the player self-reporting their age and gender, and being given Jack (boy) or Jackie (girl) based on gender choice. The opening story animation shows a house being cracked open by a growing beanstalk from a teddy bear dropped from the sky. A house floor board becomes the balance beam. A friendly creature pleads "Can you please return my teddy bear?" and by keeping the beam balanced, the beanstalk progresses upward to the goal. Jack/Jackie sees progress with the background sky changing and a progress bar with the creature noting progress between tiers. The win state is reached after Tier 6 of problems (a "bonus" Tier 7 of challenge problems is then offered for further play, an appendix of sorts). The win state is challenging: of the 240 children whose logs are analyzed in the next section, 16 (9 boys, 7 girls) saw the win state and "beat the game" by returning the teddy bear to the creature on the moon. The win state is achievable after 30-45 minutes of play and success on at least 54 balance problems (9 each in 6 tiers).

Tiers 1 through 7 get increasingly complex, starting with problems where simple mirroring works (as in Figure 1; watering two flowers into place one to the right of fulcrum will balance the two bugs located one to the left of fulcrum; the opening tutorial notes that flowers weigh the same as bugs). Later problems have pod slots which are not active and can't grow flowers, and reduced inventory so that a player will need to account for distance from the fulcrum. Specifically, Tier 1 is Siegler Rule 1: simple problems where mirroring distance works with only one piece on each side. Tiers 2 and 3 are Siegler Rule 1 and 2: matching distance and weight separately. Tier 4 combines distance and weight with more active pods (fewer blockage cues). Figure 2 shows Jackie in a no-SEL version on a Tier 4 problem: with only two waters in the inventory, the solution cannot be to mirror the four bugs. Tiers 5 and 6 address the sum of cross products (Siegler Rules 3 and 4) in increasing complexity. A study with youth ages 5-19 showed that the use of Rule 4 did not occur with children younger than 14 [3]. *Beanstalk*, targeting proportional reasoning in this problem space for children ages 5-11, must emphasize Rules 1 and 2 especially.

Figure 2. Tier 4 problem in no-SEL version (no Chicken or Crow, just general narration accessed with "?" help button).

Between tiers, players in the POHE treatments receive an in-game test sequence associated with the problems just completed. Seagulls fly onto the beam which is held up by Chicken and Crow in SEL, or plain triangles for no-SEL. They are asked in voice-overs and cloud "possibilities" to predict what will happen when the supports are removed, as shown in Figure 3.

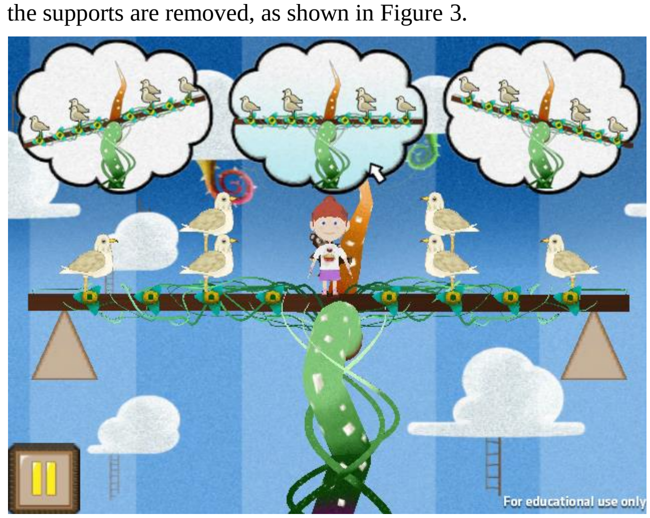

Figure 3. Predict step for POHE with no-SEL treatment.

After a prediction is made (and logged for correctness), the player sees the beam rotate or stay in place, and during the observation time is prompted to think about what happened and formulate a hypothesis. The player then is shown 3 options along with text and audio (for non-readers) to select a hypothesis. The player finishes with the seagull scene with explanation about the correctness or not of the hypothesis step, which was the second in-game test (prediction was first) during this POHE sequence. Sometimes, arrows and numerals are animated in the explanation scene. These steps of predict, observe, hypothesize, and explain utilizing talking characters Chicken, Crow, and Jack are stepped through in Figure 4. After explanation, the player sees a tier advancement scene, showing game progression up to the creature on the moon. Players in the no-POHE treatments proceed directly to the tier advancement scene directly, connecting one problem tier scene to the next.

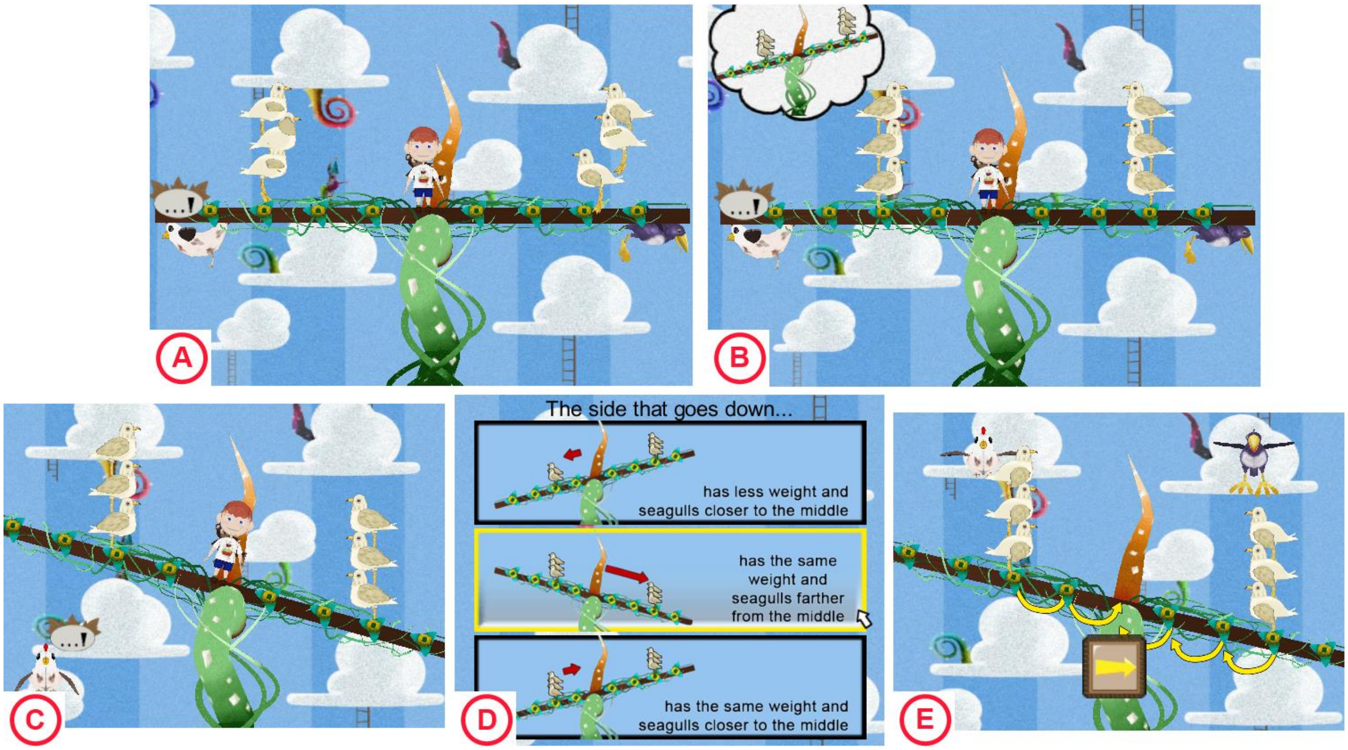

Figure 4. Sequence of scenes for POHE with SEL treatment: (A) seagulls arrive to beam held up by Chicken and Crow who chat with player; (B) each of Chicken, Jack, Crow introduce their thought on what will happen and player predicts one as correct; (C) player observe what happens when beam is free to rotate with more discussion; (D) player makes hypothesis choice based on given seagull setup; (E) player sees explanation with more Chicken/Crow dialog and counting arrows.

3. GAME LOGS

Players are randomly assigned to one of four treatments, with player data anonymized with a user ID and session ID and stored in dated log files on a central server. From Learning.com, there were 276 users, of which 36 self-reported as adults. This analysis looks only at the 240 self-reported non-adult players from Learning.com, who produced 98,126 game actions across 478 sessions. 22 of the 240 changed their mind at some point about their demographics through menu actions or across sessions. Eight times it was increasing age by 1 from an early session to a later one (perhaps a birthday took place), and only 2 times a change of gender was given. So, less than 10% of users changed their self-reported demographics, and some changes might be truthful. A spot check of sources of game traffic turned up elementary schools, indicating a likelihood of child players. While we do not know for certain that a player self-reported as a 9-year old boy is indeed 9 or a boy, we have some evidence that more than 90% stay with their selections. There were 130 girls and 110 boys, age range shown in the pie chart of Figure 5. 89% (213) were in our target demographic of ages 5-11.

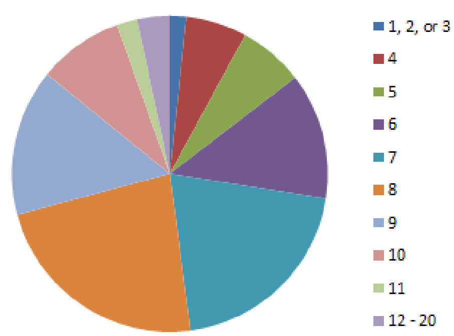

Figure 5. Distribution from 240 players' self-reported ages.

3.1 Tier Advancement

Well-designed educational games attempt to keep players in a good flow zone [2, 6]. Advance problem difficulty too gradually, and players get bored with the repetition and leave the game. Advance too quickly, and players leave out of frustration. *Beanstalk* endeavors to use adaptive learning, so that players demonstrating success advance to the next tier more quickly than struggling players: the player's demonstrated performance in the game dictates the difficulty progression pace.

The average age of the 240 players was 7.53. 40 players dropped out during the opening story scene or tutorial before Tier 1 problems began. The average age of the remaining 200 players was exactly the same: 7.53. The average age of all players reaching a given tier is given in Figure 6. As designed, tiers become more difficult. From prior research [3], we know that age is correlated with understanding of Siegler Rules which in *Beanstalk* are correlated with tiers. It is not surprising that younger players start to drop off with advancing tiers.

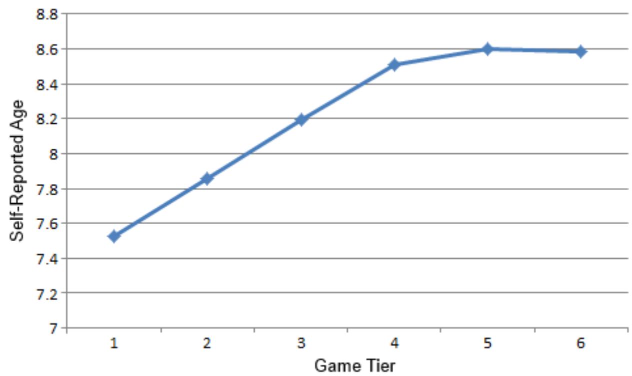

Figure 6. Average age of players starting a given tier.

Table 1 presents some overview data for the players organized by treatment. 83.6% of No-SEL players stayed to start of Tier 1 (until end of Tier 1, there is no POHE for any player and no consideration yet of the POHE treatment differences). 83% of SEL players stayed to the start of Tier 1. Hence, there was no drop-off immediately of players due to the presence or absence of the NPCs Chicken and Crow and their SEL dialog. There were no significant differences on the number of problems addressed in player sessions or player time with the game across the four tested treatments (time reported as mm:ss in the table and elsewhere).

Table 1. Player counts across treatments.

Treatment	Started Game	Started Tier 1	Started Tier 2	Retention to Tier 2	Avg. Play Time
No-SEL, No-POHE	66	57	32	48%	28:19
No-SEL, POHE	62	50	32	52%	31:47
SEL, No-POHE	51	44	29	57%	30:07
SEL, POHE	61	49	25	41%	28:34

Once Tier 2 starts, the players in No-POHE would have skipped over a "predict-observe-hypothesize-explain" opportunity and those in POHE would have seen it, i.e., from this point forward players in the 4 treatments do have 4 different experiences. When do such players quit the game? A check of the 118 players who started Tier 2 produces the chart of Figure 7. The game version that includes both socio-emotional learning and in-game tests causes the children to stop playing the game sooner. That pattern is even evident in drop-off from Tier 1 to Tier 2 in Table 1: SEL, POHE retains the fewest percentage-wise, perhaps because they are dropping off when seeing the chatty Crow and Chicken bolstered POHE dialog sequence after Tier 1 finishes but before Tier 2 starts. The evidence of Table 1 and Figure 7 is that SEL combined with POHE is not motivating children to stay with the game and deal with more complex problems in later tiers.

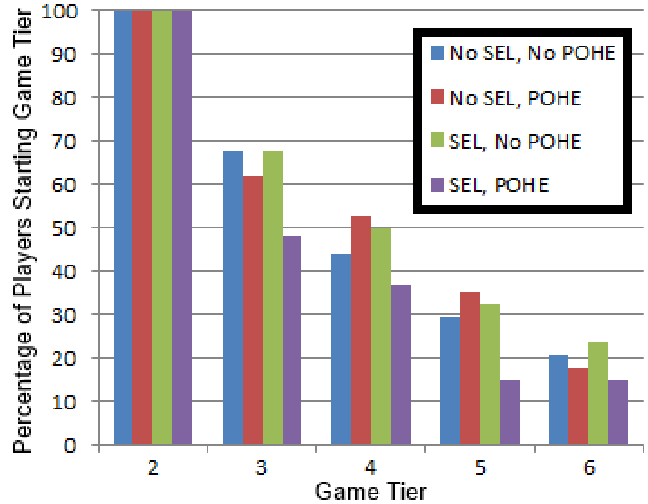

Figure 7. Percentage of players starting tier, looking at the 118 that started Tier 2 (and hence had no-POHE vs. POHE).

Figure 8 presents the accuracy per tier per treatment. An accuracy of 1.0 indicates a perfect 9 correct of 9 required correct responses (with the needed number of correct in a row to advance for adaptive learning occurring as well, since there were no errors), for all of the players in the treatment. Each tier after Tier 3 is progressively more difficult. The spike at Tier 3 reflects the fact, that while there are more complex arrangements of objects, simple mirroring will still solve these problems, so the children have had a lot of practice. The drop at Tier 4 shows a new rule is needed to solve the problems. The Siegler task space is interesting, in that at early levels, 5 year olds can succeed by mirroring, but at the hardest levels teens can stumble as they must use the sum of cross product rule if presented with a restricted inventory.

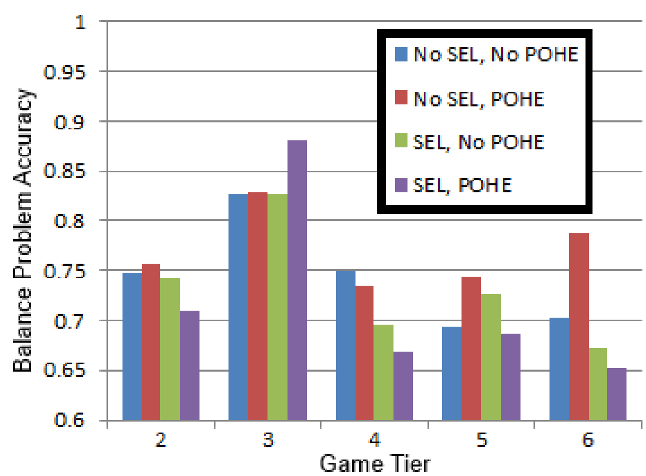

Figure 8. Problem accuracy across *Beanstalk* game tiers.

Figure 8's chart also shows that the children who persist do as well on the difficult levels as they do on the easy levels. Shute and Ventura argue that in appropriately designed educational games, completion of tasks in the game is evidence of learning [7]. With nearly identical accuracy on all tiers, the players that persist are showing consistent learning of progressively more complex rules. That is, completing Tier 3 but not Tier 4 shows the simplest understanding of the problem: a greater weight will cause that side of the beam to go down. In Tier 4 they demonstrate the knowledge that, given equal weights, the distance from the fulcrum determines balance. In Tiers 5 and 6, the sum of cross product rule is needed: the sum of the weights times the distances on each side must be equal to balance the beam. Looking at Figure 8 for the difficult Tiers 4, 5, and 6, and it is clear that the in-game testing of POHE when offered in a straightforward manner (No-SEL, as in Figure 3) does best. "No-SEL, POHE" performs well on the difficult tiers, with players given the extra opportunity of in-game tests with some dialog (a few minutes) for reflection and feedback. "SEL, POHE" users do not fare as well, i.e., those who see POHE overlaid with helping Chicken and Crow through an argument with socio-emotional implications. The overlay is lengthier and might (1) distract the user; (2) dilute their attention; (3) bore the user.

There was nothing noticeable among the 16 (from 240) players who beat the game and restored the teddy bear to the monster. 9 were male, 7 female, with the shortest time to victory a bit over 36 minutes at 36:17. One winner reported her age as 7, seven as age 8, four as age 9, three as 10, and one as 12-20. Seven saw POHE (so nine did not), and seven experienced SEL (so nine did not).

3.2 In-Game Testing through POHE

Figure 9 shows the players' performance on the Predict activity (done first, Figure 3) and Hypothesize activity (done second, Figure 10) for the children receiving POHE. They either saw it delivered with Chicken, Crow, and SEL extra dialog, or no NPC, no SEL, and straightforward objective narration. The players receiving SEL did worse across the tiers in the in-game testing. Clearly, SEL interfered with POHE.

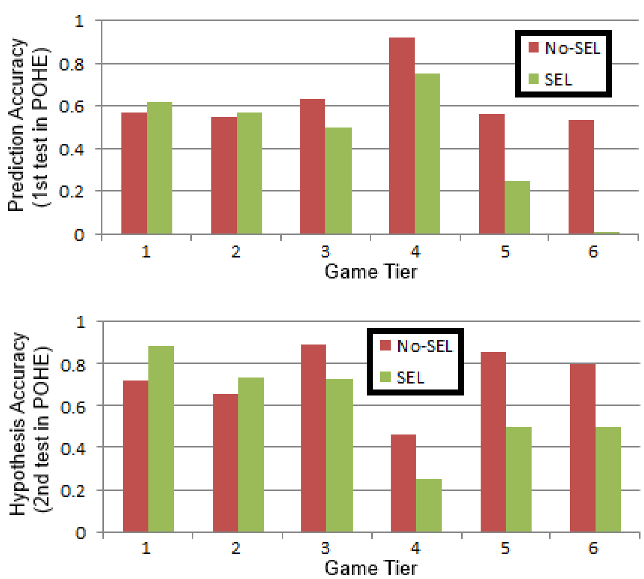

Figure 9. In-game testing accuracy across tiers.

The designed-for pedagogy was that the observation between prediction and hypothesis, coupled with discourse and the experience of the prior tiers' problem-solving, would produce a better performance on hypothesis (the second test in the POHE sequence) over prediction (the first test). This occurs for both No-SEL and SEL across all game tiers except Tier 4, i.e., for 10 of the 12 data pairs shown in Figure 9. Figure 9 shows an anomaly with Tier 4. In looking at the POHE step at that tier, the prediction is extremely simple and is shown in Figure 3: by Tier 4 the player is well-versed in "mirror produces balance" and this is a symmetric mirrored setup. Hence, the prediction is very easy and scores highly. The hypothesis for Tier 4 is shown in Figure 10. The pedagogy somehow skewed from "introduce distance and weight combinations in the simplest of ways first" into a question for which there are three correct answers, not just one, yet only the middle choice was graded as "most correct" from the set. This flaw is discovered by looking at game logs and seeing results like Figure 9 for Tier 4: clearly all three of these options shown in Figure 10 are correct and players are unjustly penalized for choosing the top or bottom option. Those options do not apply to other problems such as the answer to Figure 2, but they do apply to the set-up shown in Figure 3 and within the pictures of Figure 10. A review of the other tiers' POHE sequences show no other poorly framed hypotheses: in all other cases there is a single clear answer. Prediction (first test, Figure 3) did not suffer from wording as the player simply chose whether the beam tilted right, left, or stayed level, with the result always unambiguous.

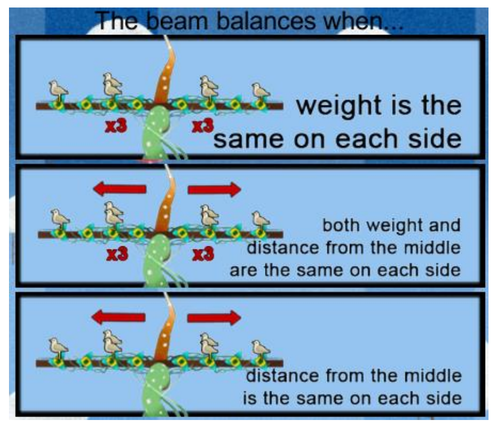

Figure 10. Hypothesis interface for POHE treatments (also showing faulty choices for Tier 4: all are correct).

4. DISCUSSION AND CONCLUSIONS

Beanstalk benefited by first iterating through designs and playtests with small groups of children in the target age range [1]. The requirements drawn from that formative work included the need to set the experience into a story, to have the story be meaningful to both boys and girls so that there is no gender bias, and to be sure the story is easily understood so that there is no cognitive overload of dealing with the educational task in the context of the story. That opening work for *Beanstalk* concluded with a test of tens of children for different versions of the game [1]. Issues were discovered and fixed, e.g., not forcing cooperation but allowing for it in a socio-emotional learning (SEL) treatment of the game. The plan was to then test via a web-based deployment. After a year in the field, with no marketing at all, there were 276 users drawn to one tested portal. By contrast, other educational game researchers have tested with tens of thousands [4], but did so benefitting from a close relationship with the game portal in which the game was featured on landing pages. In retrospect, we needed to consider marketing the game and working with the portal to promote its use, especially during a focused testing period. Here, we just put it out there to see what happened, producing enough data to get insights and further points for game correction, but not an impressive volume of users. One quick lesson from our work is the importance of marketing.

A key feature of *Beanstalk* is adaptive learning: players who fail to solve problems remain at that same difficulty tier, receiving increased amounts of scaffolding in the forms of voice-over suggestions and eventually ghosting in solution states. Guess-and-check strategies whereby a child throws anything at the beam and waits for more revealing hints are discouraged by the game tier not advancing until the player demonstrates repeated success without guess-and-check. The level design for *Beanstalk* does progress forward through Siegler Rules, which makes succeeding tiers more difficult. Game log data confirms that player ages for those starting a tier increases by tier. Player performance does not drop off, in part due to the adaptive strategy: they only advanced to a Tier N if they demonstrated competency with Tier N-1.

Socio-emotional learning aspects of persistence (giving up too early), asking for help, cooperation, and solving through discussion were tested, and found to have a negative effect on the experience: children in SEL treatments performed weakest on the advanced tiers' problems. When SEL was paired with a predict-observe-hypothesize-explain (POHE) in-game test sequence, children stopped play sooner, and performed weakest on the in-game prediction and hypothesis tests. SEL was present in the form of a chatty Jack/Jackie, playing the role of peer/friend to the player, Chicken (eager to help, positive and excitable), and Crow (also likes to help, but preens when correct and likes to take credit for player's correct activity; sharper than Chicken but not as eager). The no-SEL treatments were less chatty with only a generic narrator giving audio help when requested and stepping through the POHE sequence.

Introducing additional storylines with Chicken and Crow and adding complexity to the interface with these characters hurt the experience over time. Initially, in Tier 1, players persisted in the game the same between no-SEL and SEL. By Tier 2, after POHE was experienced, players in SEL with POHE dropped off faster than players in other treatments. A story's set-up of the game can work, but at some point that set-up needs to recede into the background. The primary game mechanic in *Beanstalk* is watering parts of the beam to put it into balance. If instead, the interactions with Chicken and Crow become too dominant, the game reduces to linear exchanges of dialog with these NPCs rather than working through beam balance issues. Children, by leaving the game sooner and performing within it more poorly, judged the SEL with POHE as the weakest variant.

Prior research with the balance scale task showed an age dependency: Siegler Rule 4 may not be as accessible to young children [3, 9]. *Beanstalk* confirmed an age correlation with difficulty tiers. *Beanstalk* did not push the adaptivity as much as it could have. Tiers remained tied to Siegler Rules, e.g., Tier 6 will show Siegler Rule 4 problems. That locks in the flow diagram of play for all users such that all users will need to be able to address Rule 4 problems in order to pass Tier 6. Given this particular task, a more flexible approach capable of addressing the wide age range of children ages 5 to 11 would be to set the Rule being addressed in a tier based on performance of the prior tiers. A player succeeding well in the game through a few tiers can get accelerated problem difficulties, placing him or her in a good flow zone of not too easy, not too frustrating. Perhaps such a child is age 11 and has the needed skills to succeed. A player failing regularly in a tier but getting by just enough to progress to the next can be presented with a milder increment in problem difficulty. This balance task is well suited to such increments of difficulty. Perhaps this child is age 5 and this is the first experience with playing with such a problem. Ideally, that same child can return to the game later and play again, and now experience a more steeped progression through difficulty because now he or she can get past easy problems with higher demonstrated success. Rather than tie Tiers to explicit rules, the next iteration of the balance game being worked on by the developers is leaving the full tier progression subject to adaptive learning. By demonstrating success with in-game problems, future problem difficulty is accelerated. If players start to falter, problem difficulty is relaxed.

Much can be learned by mining educational game logs. For *Beanstalk*, it uncovered a pedagogical issue in a tier's in-game assessment, and differences between four fielded systems of a game. Serious game analytics, i.e., inspecting player experience through the log trails they leave in games, allows for data to be collected to inform adaptive learning strategies. An opening challenge is to field a game that will be picked up and played by the target demographic: recreational games have set the bar high, so consideration of good game design principles needs to be followed to make the educational game competitive for the players' attention. The data reported here will enable the next development iteration of the game to be better suited to delivering a stronger pedagogical experience with the balance scale task.

5. ACKNOWLEDGMENTS

This work was funded by DARPA's ENGAGE program. The ENGAGE researchers at Carnegie Mellon HCII were an integral part of *Beanstalk* design and evaluation. They include Vincent Aleven, Ken Koedinger, Steven Dow, Carolyn Rosé, Jonathan Sewall, Mitra Fathollahpour, Chris Reid, Julia Brynn Flynn, Amos Glenn, and Erik Harpstead.

6. REFERENCES

[1] Christel, M., et al. 2013. Beanstalk: A Unity Game Addressing Balance Principles, Socio-Emotional Learning and Scientific Inquiry. In Proc. Int'l Games Innovation Conf. IEEE, NY, 36-39. DOI: 10.1109/IGIC.2013.6659126.

[2] Csikszentmihalyi, M. 1990. *Flow: The Psychology of Optimal Experience.* Harper and Row, New York, NY.

[3] Jansen, B. R. J., and van der Maas, H. L. J. 2002. The Development of Children's Rule Use on the Balance Scale Task. J. Experimental Child Psychology, 81, 383-416.

[4] Lomas, D., Patel, K., Forlizzi, J. L., and Koedinger, K. R. 2013. Optimizing challenge in an educational game using large-scale design experiments. In *Proc. Conf. on Human Factors in Computing Systems (CHI '13)*. ACM, New York, NY, USA, 89-98.

[5] National Research Council. 2012. *NRC Committee on a Conceptual Framework for New K-12 Science Education Standards, A Framework for K-12 Science Education: Practices, Crosscutting Concepts, and Core Ideas.* National Academies Press, Washington, DC.

[6] Schell, J. 2008. *The Art of Game Design: A Book of Lenses*. Morgan Kaufmann, Burlington, MA.

[7] Shute, V. J., and Ventura, M. 2013. *Stealth Assessment: Measuring and Supporting Learning in Video Games*. MIT Press, Cambridge, MA.

[8] Siegler, R. S. 1976. Three aspects of cognitive development. *Cognitive Psychology* 8 (1976), 481-520.

[9] Siegler, R. S. and Chen, Z. 2002. Development of Rules and Strategies: Balancing the Old and the New. *J. Experimental Child Psychology* 81 (2002), 446-457.