Incentives in educational games: A multilevel analysis of their impact on elementary students' engagement and learning

Michael Filsecker, Duisburg-Essen University, Forthausweg 2, 47058 Duisburg, Germany, michael.filsecker@uni-due.de
Daniel Thomas Hickey, Indiana University, 201 North Rose Avenue, Suite 4000
Bloomington, Indiana 47405-1006, dthickey@indiana.edu

Abstract: The effects of incentives on engagement and learning were analyzed at multiple levels in an immersive videogame for elementary science. One group of fifth-graders was offered incentives and another group was not offered incentives. The feedback afforded by the videogame was expected to mitigate predicted negative effects of incentives. No significant motivational effects of incentives were found across engagement levels: immediate (engagement with resources), close (participation in drafting in-game reports), proximal (self-reported situational interest) or distal (gains in self-reported personal interest). Nearly all of the differences that were found favored the incentive condition. Students in the incentive condition showed significantly larger gains in conceptual understanding (proximal) and non-significantly larger gains in achievement (distal). These results suggest that the predicted negative consequences of extrinsic incentives may be addressed or even reversed in this new generation of learning environments, and point to value for a multi-level model of assessment and engagement.

While most commercial videogames offer players some form of *incentives* (such as points or "levels") to motivate their progress, incentives remain controversial in education. Cognitive theorists assume that incentives undermine intrinsic motivation and subsequent engagement via the *overjustificatio*n effect (Deci, Ryan, & Koestner, 2001, Lepper, Greene, & Nisbett, 1973). This occurs when an extrinsic incentive is introduced for activity which was previously intrinsically interesting. After the introduction of the incentive (e.g., a prize or a certificate) the individual subsequently attributes the basis for the activity to the extrinsic reward. Hundreds of studies have shown that "extrinsic" incentives direct attention away from intrinsically motivated learning, leading to diminished engagement once incentives are no longer offered (Tang & Hall, 1995). Reflecting the antithetical relationship between cognitive and behavioral theories of motivation, analyses of the same body of studies by behaviorally-oriented theorists support the conclusion that the negative consequences of incentives are limited to specific easily-avoided situations (Cameron & Pierce, 1994).

Sociocultural Perspectives on Incentives

Newer sociocultural theories of knowing and learning offer a different way of thinking about incentives and motivation that might move this debate forward. In their groundbreaking paper on *cognitive apprenticeship,* Collins, Brown and Newman (1989) suggested that the corrosive educational effects of competition (which is typically fostered by incentives) may be more the results of impoverished learning environments that lacked opportunities to improve and the formative feedback needed to do so. Most of the prior studies of incentives were conducted in highly structured laboratory settings or very traditional classrooms. This suggests that the newest generation of educational videogame incentives might have positive consequences that outweigh or even eliminate any negative consequences. Furthermore, the rich interactive narratives in the latest generation of immersive videogames and the participatory culture of many networked learning environments might counter or even reverse the overjustification effect via what Gresalfi, et al. (2009) called *consequential engagement*.

The meaning of educational engagement is bound to views of learning. Prior scholars have advanced notions such as *mindfulness* (Salomon & Globerson, 1987), *intentional learning* (Bereiter & Scardamalia, 1989) and *committed learning* (diSessa, 2000). As Dewey put it a century ago "…the educational significance of effort, its value for an educative growth, resides in its connection with a stimulation of greater *thoughtfulness* not in the greater strain it imposes" (Dewey, 1913, p. 58). Sociocultural approaches highlight Dewey's thoughtfulness as the process by which students engage in an activity, interact with each other and use resources and tools purposefully. Engel and Conant's (2002) notion of *productive disciplinary engagement* highlights (a) the number of students making substantive disciplinary contributions, (b) the number of disciplinary contributions made in coordination with each other, (c) students attending to each other and making emotional displays, and (d) students spontaneously reengaging. In this characterization, the role of discourse is key to supporting any claim concerning engagement.

Multi-Level Assessment Model

The model in this study emerged in prior design-based research involving *GenScope* program for learning genetics. The first cycle analyzed and fostered learning at three levels, including the *immediate-level* enactment of the GenScope activities, *close-level* informal assessments and "feedback conversations," and a *proximal-level* performance assessment (Hickey, Kindfield, Horwitz, & Christie, 2003). The second cycle added a fourth *distal-level* external test that documented significant achievement gains on targeted standards without resorting to expository instruction (Hickey & Zuiker, 2012). Subsequently in the context of immersive games, this study explored the usefulness of the model for providing valid inferences of the translation of the intense engagement with videogames to academic subject matter (Roschelle, Kaput, & Stroup, 2000). The difficulty of such translations lies, in part, on the unique affordances of educational games (i.e., formative feedback and numerous low-stakes opportunities to improve). While the formative assessment functions of these features enhance learning, they can compromise evidential validity of assessments used to examine engagement and learning in videogames. This study assumes that doing so calls for assessments along different "levels" of learning outcomes (Ruiz-Primo, Shavelson, Hamilton, & Klein, 2002). The current study assumes that using different learning outcomes across levels means that formative feedback at one level does not directly coach or prepare students for the outcomes at the next level. This provides a tractable way of controlling for the construct-irrelevant variance (Messick, 1994) that occurs when students are given feedback for solving problems that are similar to the problems that appear on an assessment (Hickey & Anderson, 2007; Hickey et al., 2006). This maximizes consequential validity (i.e., the formative function of assessment) at one level while preserving evidential validity at the next level (the summative function). Doing so across three or more levels promises to overcome the complexities of assessing learning outcomes from educational games (e.g., the concerns over assessment sensitivity raised by Annetta, Minogue, Holmes, & Cheng, 2009, p. 79).

This study extended the multi-level assessment model as it had emerged in design studies of Quest Atlantis's *Taiga* ecology game (Barab, et al., 2011) to the study of incentives and their impact on engagement. Learning was conceptualized in terms of the four levels shown in Table 1. Generally speaking, these levels were pragmatically informed by the three "grand theories" of learning outlined in Greeno, Collins, & Resnick (1996). First, a situative/sociocultural perspective was used to conceptualize (1) the immediate-level enactment of sequences of inquiry-oriented game activities and (2) close-level participation among the player, teacher, and non-player characters in writing and revising written "quests" after those activities. The model then uses a cognitive/rationalist perspective to frame learning in terms of (3) proximal-level conceptual understanding assessed with a curriculum-oriented performance assessment. Finally, the model uses a more behavioral/associationist perspective to frame learning in terms of (4) distal-level achievement measured with a multiple-choice test. This means that the collected evidence of close, proximal, and distal learning (a) were increasingly removed from the enactment of the Taiga inquiry activities, (b) were increasingly oriented towards a broader curricular scope, and (c) used increasingly abstract representations of the targeted knowledge.

Table 1: Multilevel assessment model

LEVEL (Orientation)	Assessment Format	Learning Outcome	Relationship to Curriculum	Feedback Timescale	Primary Summative Functions	Primary Formative Functions
IMMEDIATE (Action/ procedures)	Analysis of log-files, "live" discourse and social interaction	Enactment of actions and procedures, communal discourse	Same content and context	Minutes	Actual enactment of QA activities	Foster discourse and intentional learning
CLOSE (Activity)	Analysis of content of quest submissions	Interactive discourse & intentional learning	Same content and context	Hours-Days	Enactment of preceding QA activities	Foster individual understanding
PROXIMAL (Curriculum)	Open-ended problem solving assessment	Individual understanding of targeted concepts	Same content in similar context	Weeks-Months	Understanding of concepts targeted	Refine curriculum and compare versions
DISTAL (Standards)	Externally-	Aggregated	Same &	Months-	Measure	Inform broad

	developed test items	achievement of targeted standards	different content in new contexts	Years	impact on achievement	audience of curricular impact

Each level of analysis has potential summative and formative functions. For example, the close-level analysis of the questing activity has a summative function relative to the game activities but has a formative function relative to the understandings that individuals take away from that interactive writing. Aligning learning across levels reveals the presence or absence of "echoes" across levels. This distinguishes the actual consequences of design features from random variation. For example, when marginally significant distal outcomes are correlated with larger, statistically significant proximal outcomes, designers should be less inclined to dismiss the distal outcomes as having occurred by chance (see also Schaffer & Serlin, 2004). This is useful when attempting to ensure that refinements to specific curricula are consistently impacting learning on distal measures, even when working with small numbers of learners typical of early-stage design studies.

This study attempted to extend the multi-level assessment design model to the issue of incentives by building on emerging situative/participatory approaches to motivation (Greeno et al., 1998; Hickey, 2003). At the close level, we examined students' written quests as evidence of their success while participating in the interactive practice of drafting a quest. While this method lacked the attention to context or a more interpretive discourse analysis, it seemed the most appropriate way of capturing participation at this level in a way that could be interpreted in the quasi-experimental comparison. At the proximal level, we examined individual players' self-reported motivational states during that same quest. At the distal level, we examined players' more enduring self-reported personal interest towards the kinds of problems they were solving in the game. This relationship between assessment and motivation are explored in more detail in Hickey & Schaffer (2006).

Methods

This study was the third in a series of annual design studies of the 15-hour Taiga curriculum with the same teacher and population of students. In the previous year, new formative feedback resources (e.g., teacher rubrics for reviewing and giving students feedback) and new cut-scenes with useful information were added to help students complete the crucial second quest. These new resources substantially increased gains in understanding and achievement, but only for those students who accessed them (Hickey, Ingram-Goble, & Jameson, 2009). This suggested that using strategies to motivate students to access more of the resources and do so more meaningfully should further enhance learning outcomes. Incentives seem to be a promising strategy.

A quasi-experimental design was conducted to examine the effect of providing incentives on students' engagement and learning science. For two of the classrooms in this study, the teacher's acceptance of a written quest at one of three increasingly accomplished levels (proficient, expert, or wise) was rewarded with a corresponding badge that players could affix to their in-game virtual avatar (Figure 1a). Additionally, students in this Public Recognition (PR) condition were invited to move a paper version of their avatar up and across a physical "leader board" that was prominently placed in the room (Figure 1b). In two other classrooms taught by the same teacher in the same semester, students in the Non Public Recognition (NPR) condition were not offered badges or a ready means to communicate their level of progress to the other students and in-game information on incentives was replaced by messages encouraging players to work hard to save the park and become more capable apprentices (Lepper & Malone, 1987). The study tested the following hypotheses:

Hypothesis 1**:** Students in the PR condition will engage more deeply in the process of drafting and revising their quests, use more relevant scientific formalisms, and use those formalisms more correctly than students in the NPR condition.

Hypothesis 2**:** Students in the PR condition will exhibit significantly larger gains in conceptual understanding of the targeted science concepts and achievement of the targeted science standards than students in the NPR condition.

Hypothesis 3: There will be no difference between the PR and NPR conditions in self-reported intrinsic motivation during the second quest, and no differences in impact of the game on personal interest in learning to solve these types of scientific problems.

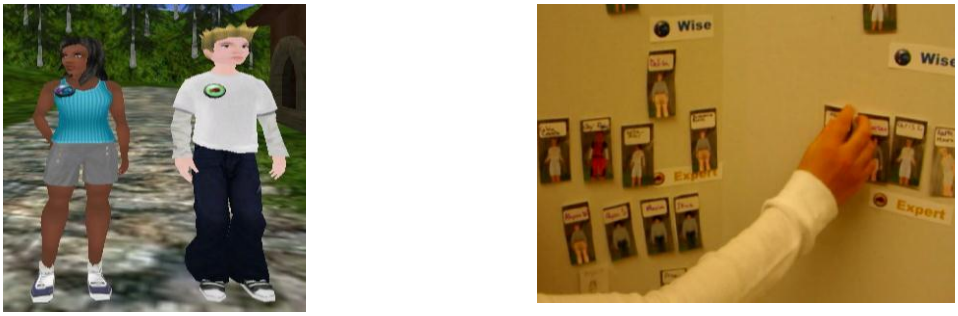

Figure 1a and 1b: Avatar Badges and Leader Board Displaying Levels.

Participants and Materials

This research was conducted at a public elementary school in a medium-sized city in the Midwestern US. As is typical of university communities, the students were predominantly Euro American and most came from well-educated professional families. In this study, average grades from prior work were used to identify pairs of similar achieving classes, and one class in each pair was assigned to the Public Recognition (PR) and the Non Public Recognition (NPR) condition. Consent to participate in the study was obtained from almost every student, resulting in 106 participants (56 females and 60 males).

Instrumentation and Procedures

Learning and engagement were assessed simultaneously at the immediate and close levels and separately at the proximal and distal levels. At the immediate level we analyzed the number of screens of formative feedback that students accessed by accessing the log files generated during gameplay as in the previous studies (Hickey et al., 2009). This reflected our tentative assumption that choosing more pages represented more intentional engagement in the structured discourse of the revision process. To assess learning and engagement at the close level, we analyzed the quality of the initial and final submissions of crucial Quest 2 (scored by researchers) using a 14-point scale rubric which assigned six points for summarizing the water quality indicators, four points for explaining what the processes were (i.e., erosion and eutrophication), and four points for describing the dynamic relationship between indicators and processes.

While this rubric could capture the students' right or wrong answers to Quest 2, it could not capture the meaningful appropriation of concepts in the domain discourse. For example, one student could say *dirt from Site B got into the river*, while another one could say *the sediment from Site B is eroded into the river*. By using the 14-points rubric, both students would have earned one point, without distinguishing the nuances such as the difference between *dirt* and *sediment* and between *got into* and *eroded*. In a sense we were aiming at the *disciplinary* engagement pointed out by Engle and Conant (2002). Therefore, we quantified the verbal data (Chi, 1997) to capture this domain-specific or *disciplinary* discourse around students' Quest 2 submissions (n=106). Initial and final submissions in Quest 2 were coded in terms of the meaningful appropriation of nine relevant scientific concepts. The text of the submissions of all students (n=106) was coded using the NVivo qualitative analysis software program. We were interested in capturing students' engagement with the content in a progressive knowledgeable way as a result of the incentive manipulation, instead of students' actual representation of knowledge (e.g., Chi, 1997), scientific argumentation (e.g., Kelly, Drucker & Chen, 1998; McNeil et al., 2007) or knowledge construction (e.g., Weinberg & Fischer, 2006).

Table 2: Examples of a few scientific concepts coded as accurate, non-accurate or other.

	Examples of evidence for coding		
Scientific concepts	Accurately	Non-accurately	Other
Dissolved Oxygen	Warm temperature takes out the DO in the water so the fish suffocate… (Student ID 118412)	The Tempature is effected by Do level in the river (Student ID 118517)	There is too much of every thing except DO and it is way to hot. (Student ID 118306)
Turbidity	Turbidity is caused some by erosion. (Student ID 118221)	without turbidity the sun will get through the water and then the plants can't grow. (Student ID 118509)	At site A and C, the Turbidity is in between. (Student ID 118504)

Table 2 provides examples of actual students' answers (misspellings in original responses) and the category in which they were coded. The meaningful appropriation of the concepts has to do with (1) identifying the right level of the indicators displayed in the charts, (2) the concept being used to establish a valid relationship with other concepts, (3) relating the concept to a relevant activity or event, and (4) identifying the concept as being the cause/effect of another concept or associated activity or event. A non-meaningful appropriation situation has to do with (1) establishing the concepts in an invalid relationship with one another (2) the concept being used to explain the wrong ecological process (e.g., erosion or eutrophication), or (3) the concept being used incorrectly as a cause or effect of an event or another concept. The category "other" was used when the student's response was too ambiguous to discern the appropriate category.

To examine engagement at the proximal level, we developed a scale to assess players' situational motivation regarding the Quest 2 activity. The scale consisted of 4 or 5 Likert-type items (strongly disagree, disagree, neutral, agree, or strongly agree) for each of the following subscales of the motivational states that prior research has shown to be diminished by incentives: *interest* in the activity, *value* for completing the activity, *perceived competence* during the activity, and *effort* completing the activity. So long as the individual scores for each set of items are internally reliable, scores on each scale are presumed to be indicative of various aspects of students' cognitive engagement during the tasks (see Fredricks, Blumenfeld, & Paris, 2004). Once their Quest 2 submission was accepted, students completed the brief survey. The survey asked students, "How did you feel while completing Quest 2?" The survey also encouraged students to respond honestly and assured students that their responses were confidential.

To examine engagement at the distal level, we measured changes in personal interest in solving the types of problems students were learning to solve in Taiga. One of the main concerns with incentives is that they may supplant existing intrinsic motivation towards activities with the extrinsic motivation offered by the incentive - the "overjustification effect" (Lepper et al., 1973). Hundreds of prior studies in laboratories or traditional classrooms showed that extrinsic incentives lead to decreased free choice engagement in the incentivized activity. Many of those studies also examined self-reported interest in the activities (and sometimes instead of) free choice engagement. To this end, we measured students' self-reported personal interest in the three types of problems that they were learning to solve in Taiga: *water ecology* problems, complex scientific problems, and controversial socio-scientific problems. An 18-item survey was created consisting of six Likert-scale items for each type of problem and was administered before and after students played the game.

To examine learning gains at the proximal level, we used the *Lee River* performance assessment developed in the prior design cycles. The assessment was "curriculum-oriented" in that it asked students to solve similar problems as in Taiga but in a somewhat different context. The assessment had been created alongside extensive refinements to Taiga the previous year and was designed to be highly sensitive to different enactments of the curriculum. It involved another fictional watershed and a range of stakeholders who had similar (but not identical) effects on the ecosystem. For example, both Taiga and Lee River involve stakeholders with different land use practices who are arranged along a river. The stakeholders from both scenarios impact their ecosystems by doing things that cause erosion and eutrophication; however, erosion is caused by loggers in Taiga and by construction in Lee River. To capture a range of understanding at the pretest and the posttest, the items covered included a broad range of difficulty. It included several multi-part items that started out with simple tasks that most students would be able to answer without instruction, and proceeded to a few complex items that focused on the nuances of scientific hypotheses, the relationship between social issues and scientific inquiry, and the relationship between water quality indicators such as dissolved oxygen and processes like eutrophication. A 21-point scoring rubric was used to score completed assessments, with a subset of assessments scored by two researchers to establish reliability.

To examine learning gains at the distal level we used the same 20-item test that had been created the previous year by random sampling from pools of items aligned to the four targeted content standards, but independent of the Taiga curriculum. Such standards-oriented tests are necessary to support claims of impact on externally-developed achievement measures and to compare the impact of different curricula that target those standards. Such tests are not particularly sensitive to specific interventions and represent a relatively ambitious target for innovative curricula like Taiga.

Results and Findings

For engagement and learning at the immediate level, analysis of the log files found no significant difference in the number of feedback pages accessed for the PR (M=8.69, SD=6.91) and the NPR (M=9.24, SD=5.98) conditions [Mann–Whitney U =1285, n_1=51, n_2=55, p=.452]. At the close level improvement scores for the initial and final Quest 2 submissions (using a 14-point scale; inter-rater reliability = .85) did not reach statistical significance between conditions [Mann–Whitney U =1099, n_1=47, n_2=51, p=.475]. In addition, a one-way MANOVA was conducted to compare the effects between conditions on the meaningful appropriation of the scientific concepts as enlisted during the drafting of Quest 2. The analysis of the coded initial and final submissions revealed higher levels in the PR condition, but the difference did not reach statistical significance

[Wilks' Lambda =.973, $F(1,102)=2.797$, $p=.097$] Therefore, strictly speaking, we found no evidence of negative consequences of incentives engagement in the written discourse around Quest 2.

Concerning proximal engagement, all four self-reported assessments of situational motivation during Quest 2 revealed high internal reliability (all alphas over .85). This was crucial, since unreliable measures could have masked consequences of incentives in random variance. A one-way between subjects ANOVA was conducted to compare the effects of the incentive and non-incentive conditions on perceived interest, value, competence, and effort. There was no significant effect on any of the variables [$F(1,106)=.826$, $p= .366$; $F(1,106)=.051$, $p =.821$; $F(1,106)=.467$, $p=.496$; $F(1,106)=.321$, $p=.575$, respectively]. While none of the four differences reached statistical significance, the fact that slightly higher scores were observed for all four of the aspects in the PR condition argues strongly against the predicted negative consequences from the incentives.

For distal engagement the scales of interest in solving the three different types of problems showed acceptable levels of reliability (alphas over .80) at both administrations. A one-way repeated measures ANOVA was conducted to compare the effects of incentives on three indices of interest. None of the tests yielded significant difference between conditions [Wilks' Lambda =.99, $F(1,102)=.442$, $p=.508$;Wilks' Lambda =.99, $F(1,101)=.703$, $p=.404$;Wilks' Lambda =.99, $F(1,101)=1{,}026$, $p=.314$], supporting our initial hypothesis that the "overjustification" is unlikely to occur in contexts such as QA. These results suggest that the introduction of incentives in this environment did not undermine personal interest (or presumably subsequent free-choice engagement) in these times of scientific investigations.

For proximal learning, a one-way repeated measures ANOVA tested the effects of incentives on students' gains in conceptual understanding. Students in the PR condition gained significantly higher levels of understanding than students in the NPR condition [Wilks' Lambda =.946, $F(1,99)=5.6$ $p= .02$]. As shown in Figure 2, this represented the difference between 1.4 and 1.1 SD gain, given the pooled standard deviations across the score points. Importantly, the differences in gains between the two classes in each condition were not statistically significant ($F < 1$). Thus, the students in the incentive condition developed significantly greater understandings of the concepts, topics and processes associated with solving scientific and socio-scientific problems involving water quality.

For distal learning, the achievement tests revealed strong internal consistency, and showed that students in the PR condition gained 5.44 points compared to 4.02 points for the other students. Given the variance within the scores, this was a difference between gains of 1.1 and 0.8 SD. A one-way repeated measures ANOVA revealed that this difference in gains did not reach conventional criteria for statistical significance [Wilks' Lambda =.972, $F(1,114)=3.234$, $p=.075$, gains between classes within groups was again $F < 1$]. However, such a gain seems highly unlikely to have occurred by chance given the corresponding significant difference in gains in proximal understanding. This is an example of the aforementioned "echo" and an illustration of the advantage of assessing learning outcomes across multiple increasingly formal levels.

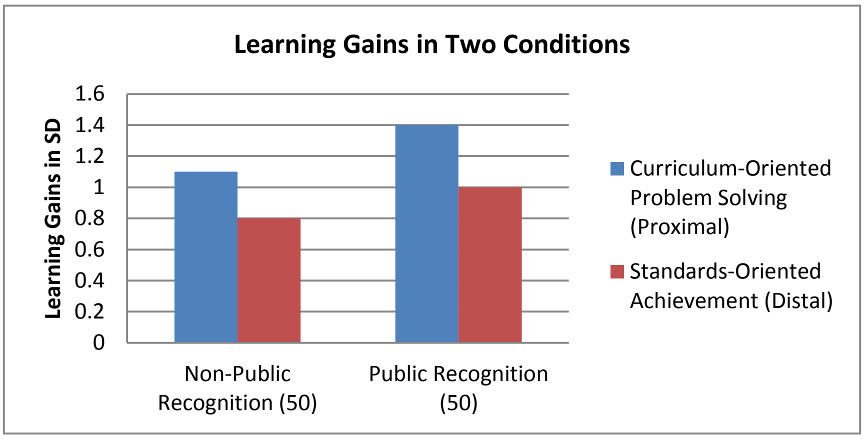

Figure 2: Proximal and Distal Learning Outcomes
by Implementations and Conditions (* = p<.05; *** = p<.01).

In summary, the incentives as enacted in this study were not shown to motivate students' engagement with the learning activities such as drafting and revising Quest 2 and using the resources embedded in the game. Therefore, Hypothesis 1 was not supported. However, results showed a significant larger gain in understanding of ecological processes (proximal), and a non-significant differential gain in achievement (distal) both in favor of the PR group. Therefore, Hypothesis 2 was partially supported. Finally, examination of engagement at the three levels failed to uncover any of the potential negative consequences of the incentives, supporting our third hypothesis.

Implications and Significance

These findings lend initial support to the argument advanced by Collins, Brown, and Newman (1989) that the negative consequences of competition may be more indicative of impoverished learning environments and the lack of feedback and opportunity to improve, than of a fundamental consequence of competition. Likewise, the study provides some initial empirical support for the speculations about sociocultural theories of engagement in Hickey (2003) and Hickey & Shaffer (2006). Rather than (a) using incentives haphazardly or (b) attempting to prove their fundamental impact, we believe that designers should ask about the motivational design features concerning their impact on immediate-level and close-level engagement in learning. While there are likely many ways of doing so, we believe that this more process-oriented and contextual analysis offers a helpful starting point. We also believe that this study shows some initial value in extending the multi-level model of assessment used in past studies to consider engagement and motivation as well.

Arguably, the multilevel assessment model applied in this study begins to address a core validity issue that has long plagued the assessment of individually-oriented motivational interventions (see Adelman & Taylor, 1994). Just as with our learning outcomes, our formative efforts to refine engagement at one level do not undermine the evidential validity of the engagement outcomes at the next level. In other words, there was nothing about close-level motivational intervention (i.e., incentives and competition) that might have directly encouraged students to characterize that activity as more interesting or engaging on the proximal-level survey items. At the same time, we indirectly examined the consequences of incentives and completion on student's self-reported cognition during those activities and in changes in self-reported interest towards those activities. This seems like a promising way around the obvious dilemma facing many motivational interventions: programs that focus directly on changing behavior may deliver behavioral change, but fail to impact cognition, while programs that focus directly on cognition may indeed impact cognition but fail to deliver enduring changes in behavior. Likewise, the model represents an extension and may well complement current analytical strategies based on discourse and video analysis (e.g., Azevedo, 2006; Engel & Conant, 2002) by introducing performance and achievement levels together with self-reported motivational states. In summary, while protecting the validity of outcome claims, the model also emphasizes the assessment of the process of engagement and learning encompassing the "hybrid research methodologies" characteristic of multidisciplinary fields such as CSCL (Stahl, Koschmann, & Suthers, 2006). Thus, the model provides a promising solution to the assessment of learning beyond sociocultural perspectives on teaching and learning.

Finally, the increased learning outcomes across the three design cycles demonstrates the broader value of this assessment driven multi-level model of assessment. While the present study focused on the impact of incentives, numerous other refinements had been made to the Taiga curriculum that were informed by evidence obtained at the various levels. Of course, some (but certainly not all) of the increased gains were due to teachers learning. Most innovators who have attempted to impact valid measures of external achievement know how difficult it is to obtain gains of this magnitude without resorting to expository direct instruction. In addition to offering a way forward on enduring design controversies like incentives, the multi-level model appears to be a promising way to deliver the evidence of achievement impact that is demanded by many educational stakeholders without compromising the more authentic learning supported by innovations like Quest Atlantis.

References

Annetta, L. A., Minogue, J., Holmes, S. Y., & Cheng, M. T. (2009). Investigating the impact of video games on high school students' engagement and learning about genetics. *Computers & Education, 53*(1), 74–85.

Azevedo, F. S. (2006). Personal excursions: Investigating the dynamics of student engagement. *International Journal of Computers for Mathematical Learning, 11*(1), 57–98.

Barab, S. A., Gresalfi, M., & Ingram-Goble, A. (2010). Transformational play using games to position person, content, and context. *Educational Researcher*, *39*(7), 525–536.

Bereiter, C., & Scardamalia, M. (1989). Intentional learning as a goal of instruction. In L. Resnick (Ed.), *Cognition and instruction: Issues and agendas* (pp. 361-379). Hillsdale, NJ: Lawrence Erlbaum Associates.

Cameron, J., & Pierce, W. D. (1994). Reinforcement, reward, and intrinsic motivation: A meta-analysis. *Review of Educational Research, 64*(3), 363-423.

Chi, M. T. H. (1997). Quantifying qualitative analyses of verbal data: a practical guide. *Journal of the Learning Sciences, 6*(3), 271–315

Collins, A., Brown, J. S., & Newman, S. E. (1989). Cognitive apprenticeship: Teaching the crafts of reading, writing, and mathematics. In L. B. Resnick (Ed.), *Knowing, learning, and instruction: Essays in honor of Robert Glaser* (pp. 453–494). Hillsdale, NJ: Erlbaum.

Deci, E. L., Ryan, R. M., & Koestner, R. (1999). A meta-analytic review of experiments examining the effects of extrinsic rewards on intrinsic motivation. *Psychological bulletin, 125*(6), 627.

Dewey, J. (1913) *Interest and Effort in Education*. Boston, MA: Riverside Press.

diSessa, A. A. (2000). *Changing Minds: Computers, Learning and Literacy*. Cambridge, MA: MIT Press.

Engel, R.A., & Conant, F. (2002). Guiding principles for fostering productive disciplinary engagement: Explaining an emergent argument in a community of learners classroom. *Cognition and Instruction, 20*(4), 399-483.

Fredricks, J. A., Blumenfeld, P. C., & Paris, A. H. (2004). School engagement: Potential of the concept, state of the evidence. *Review of Educational Research*, *74*(1), 59.

Greeno, J. G., Collins, A. M., & Resnick, L. B. (1996). Cognition and learning. In D. C. Berliner & R. C. Calfee (Eds.), *Handbook of educational psychology* (pp. 15-46). New York: Macmillan.

Gresalfi, M., Barab, S., & Siyahhan, S. (2009). Virtual worlds, conceptual understanding, and me: designing for consequential engagement. *On the Horizon, 17*(1), 21–34.

Hickey, D. T. (2003). Engaged participation vs. marginal non-participation: A stridently sociocultural model of achievement motivation. *Elementary School Journal, 103* (4), 401-429.

Hickey, D.T., Kindfield, A., & Horwitz, P. (2003). Integrating curriculum, instruction, assessment, and evaluation in a technology supported genetics learning environment. *American Educational Research Journal, 40*(2), 495-538.

Hickey, D. T., Zuiker, S. J., Taasoobshirazi, G., Schafer, N. J., & Michael, M. A. (2006). Three is the magic number: A design-based framework for balancing formative and summative functions of assessment. *Studies in Educational Evaluation, 32*, 180-201.

Hickey, D. T., & Anderson, K. (2007). Situative approaches to student assessment: Contextualizing evidence to transform practice. In P. Moss (Ed.), *Evidence and decision making, the 106th yearbook of the National Society for the Study of Education (Pt. I)* (pp. 264-287). Chicago: National Society for the Study of Education.

Kelly, G.J., Drucker, S. & Chen, K. (1998). Students' reasoning about electricity: combining performance assessment with argumentation analysis. *International Journal of Science Education, 20*(7) 849-871.

Lepper, M. R., Greene, D., & Nisbett, R. E. (1973). Undermining children’s intrinsic interest with extrinsic rewards: A test of the “overjustification” hypothesis. *Journal of Personality and Social Psychology, 28*(1), 129–137.

Lepper, M. R., & Malone, T. W. (1987). Intrinsic motivation and instructional effectiveness in computer-based education. *Aptitude, learning, and instruction*, *3*, 255–286.

McNeill, K.L., & Krajcik, J. (2007). Middle school students’ use of appropriate and inappropriate evidence in writing scientific explanations. In M. Lovett & P. Shah (Eds.), *Thinking with data: Proceedings of the 33rd Carnegie Symposium on Cognition* (pp. 233–265). New York: Taylor & Francis.

Messick, S. (1994). The interplay of evidence and consequences in the validation of performance assessments. *Educational Researcher, 23*(2), 13-18.

Roschelle, J., Kaput, J., & Stroup, W. (2000). SimCalc: Accelerating students' engagement with the mathematics of change. In M. J. Jacobson & R. B. Kozma (Eds.), *Innovations in science and mathematics education: Advanced designs for technologies of learning* (pp. 47-75). Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

Ruiz-Primo, M. A., Shavelson, R. J., Hamilton, L., & Klein, S. (2002). On the evaluation of systemic science education reform: Searching for instructional sensitivity. *Journal of Research in Science Teaching, 39*(5), 369–393.

Salomon, G., & Globerson, T. (1987). Skill is not enough: The role of mindfulness in learning and transfer. *International Journal of Educational Research,* 11, 623-637.

Shaffer, D. W., & Serlin, R. C. (2004). What good are statistics that don't generalize? *Educational Researcher, 33*(9), 14.

Stahl, G., Koschmann, T., & Suthers, D. (2006). Computer-supported collaborative learning: An historical perspective. In R. K. Sawyer (Ed.), *Cambridge handbook of the learning sciences* (pp. 409-426). Cambridge, UK: Cambridge University Press.

Tang, S. H., & Hall, V. C. (1995). The overjustification effect: A meta-analysis. *Applied Cognitive Psychology*, *9*(5), 365-404.

Weinberger, A., & Fischer, F. (2006). A framework to analyze argumentative knowledge construction in computer-supported collaborative learning. *Computers & Education, 46*(1), 71–95.