If the Gear Fits, Spin It!
Embodied Education and in-Game Assessments

Mina C. Johnson-Glenberg, Arizona State University, Tempe, AZ, USA & Radboud University, Nijmegen, NL

David A. Birchfield, SMALLab Learning, LLC, Los Angeles, CA, USA

Colleen Megowan-Romanowicz, Arizona State University, Sacramento, CA, USA & American Modeling Teachers Association, Sacramento, CA, USA

Erica L. Snow, Arizona State University, Tempe, AZ, USA

ABSTRACT

Two embodied gears games were created. Better learners should use fewer gear switches to reflect their knowledge. Twenty–three 7th graders, playing as dyads, used gestures to manipulate virtual gears. The Kinect sensor tracked arm-spinning movements and switched gear diameters. Knowledge tests were administered. Statistically significant knowledge gains were seen. For Game 1 (gear spun one direction), switching significantly predicted only pretest knowledge. For Game 2 (gear spun two directions) switching was also negatively correlated with both tests. For game 2, those who used fewer switches during gameplay understood the construct better scoring higher on both tests. Dyadic analyses revealed the winner used significantly fewer switches. In-process data can provide a window onto knowledge as it is being encoded. However, games should stay within the learner's ZPD, because if the game is too easy (Game 1), meaningful data may be difficult to gather. The use of in ludo data from games with high sensitivity may attenuate the need for repetitive traditional, post-intervention tests.

Keywords: *Assessment, Embodied Learning, Evidence-centered Design, Gears, Johnson-Glenberg, Videogames, Kinect, Mina Johnson, Motion Capture, SMALLab, STEM Education, Technology in Education, Virtual Reality*

INTRODUCTION

The use of games as learning tools and immersive classrooms is becoming more accepted. When a comparative class is instructed using game components versus traditional pedagogy, the game-based class is usally more engaged and reports better learning (Johnson-Glenberg, Birchfield, Koziupa, & Tolentino, 2014). Some sample domains include computer science (Papastergiou, 2009), engineering homework (Coller & Shernoff, 2009), engineering classes (Coller & Scott,

DOI: 10.4018/IJGCMS.2015100103

2009) and biological sciences (Lui et al., 2014). Coller and Scott (Coller & Scott, 2009) report that the students who were randomly assigned to their video game-based course spent approximately the same amount of time on their course work as the traditional students, but a concept mapping exercise revealed the game-based students showed deeper learning compared to the traditional class students. Interestingly, a metaanalysis from 2012 found evidence for the effects of video games on language learning, history, and outcomes from exergaming, but, found little support for the academic value of video games in science and math (Young et al., 2012). It should be noted that our field is still not adroit at differentiating between games and simulations, and better games are being created in the past five years.

The empirical study of serious games is a relatively new field, those of us who work as learning scientists/game designers need to know which components are most efficacious for learning and how these components correlate with traditional knowledge tests. For example, some of the more established award mechanisms used in entertainment games, e.g., leaderboards and badges, may not translate well to classroom environments. A recently published 6-week long study by Hanus and Fox found that the two traditional entertainment game "payoffs" of a 1) public board displaying all scores and 2) the awarding of completion badges might actually hinder learning by negatively affecting intrinsic motivation (Hanus & Fox, 2015).

Some of the established entertainment games are now adding educational components to their videogames, e.g., *Portal 1* and *2.* Early adopter teachers are optimistic about games (Miller, 2012). Marlow (Marlow, 2012) presents evidence suggests that designing and making games in the context of a well-conceived design curriculum has the potential to stimulate traditional pedagogies and foster student learning, in addition to making teaching and learning more enjoyable and meaningful. We believe more research needs to be conducted to understand which aspects of educational videogames are felicitous to learning and how adding embodiment might also affect learning during a game. In addition, the field of educational videogames needs to develop new methodologies and statistics for mining in-game player data and making sense of the information generated by the learner during gameplay. One gameplay modality that may promise enormous potential for learning is that of embodied learning.

"Embodied games" is a category of videogames that incorporates gesture into the act of learning. Our contention is that action can be used to instruct and when coupled with a game format, both gesture and gaming may have powerful effects on learning and retention. The theory is that all thought, even the most abstract, is derived from an original physical embodiment. Hauk et al. describe fMRI experiments that demonstrate that when reading words related to action, areas in the brain are activated in a somatotopic manner. For example, reading "lick" activates motor areas that control the mouth, whereas reading "pick" activates areas that control the hand. This activation is part of a network representing 'meaning' or semantics. The study was done on adults and thus demonstrates that the mappings do not fade once stable comprehension is attained, that is, motor codes are still activated during linguistic comprehension long after the meaning has become stable (Hauk, Johnsrude, & Pulvermüller, 2004). If cognition and the body are deeply and irrevocably connected, then perhaps all cognition is embodied (Glenberg, Witt, & Metcalfe, 2013; Wilson, 2003). If this is true, then it seems prudent to design learning games that take advantage of how the body moves to reify concepts to be learned. That is, education could benefit from having more body-based, gesture-oriented games. At the same time theories of embodiment in education are coming of age (Lindgren & Johnson-Glenberg, 2013).

Games for education. Creating educational games that are also embodied may be especially useful for topics that have traditionally been "tough to teach". We focus on science for several reasons. First, it is a closed problem space, there is usually one answer. Thus, it is a 'cleaner' less ambiguous space to design for. Second, we hold that computers, and the rich animations they afford, should be used to facilitate comprehension of content that is not readily apparent to human perceptual systems. This means making the macro (e.g., astronomy) and the micro (e.g., bacteria, or the genome) accessible. Third, we care deeply about an informed citizenry and are concerned that too many students are leaving the science pipeline around middle school. Games may be a way to engage and retain our young men and women in the sciences. Games offer students an opportunity for stepping out of their usual identities and trying on new ones.

When a learner is able to take on the "identity" of an expert (Shaffer et al., 2009), s/he can begin to feel and react like an expert. If the expert is a scientist then the learner might show outcomes like volunteering more answers in class, mentoring others, or spontaneously writing a letter to the Mayor about zoning laws learned in the game (Squire & Klopfer, 2007). Gestures may facilitate novices on their journey to becoming experts. By definition a novice is someone whose knowledge is in parts. The "knowledge in pieces" paradigm (DiSessa, 1988), posits that learners as novices do not yet hold a coherent whole model, they perhas have pieces that do not fit together. They understand that gears spin and can aid in "work", but do not understand how the input diameter of a gear train affects the work performed. We believe that by adding the extra modality of gesture in engaging embodied games, we may aid in the creation of coherent knowledge structures for these novices. (Obviously, experts are assumed to have large, coherent knowledge structures.)

Games for change. Several videogames are currently being built for the domain of science education that specifically attempt to leverage the idea of knowledge in pieces (Clark et al., 2014). The new generation of teachers reports a willingness to use games in the classroom, and the education market is concurrently seeing more cost-effective methods for motion capture. Our lab believes that merging of these phenomena (intriguing science games) with affordable motion capture technologies (e.g., the Microsoft *Kinect* sensor, embedded 3D laptop cameras *RealSense,* laser scanners, etc.) may result in game-changer tools for education and other fields (Johnson-Glenberg, Birchfield, et al., 2014; Johnson-Glenberg & Hekler, 2013; Lindgren & Moshell, 2011). We believe educational game designers need to think about how to mesh learning via embodied perceptual symbols (Barsalou, 2008) with gestures while taking into account the learning profile or the KSAO's (knowledge, skills, abilities, other) that students bring to the task. In this manuscript we focus on the knowledge component of KSAO. The games are designed to address knowledge change, but to do that one must first measure prior knowledge. Prior knowledge has traditionally been gathered via computer-based or paper and pencil tests (this study's test is included in the Appendix). The primary goal was to ascertain how the in-process, or in ludo, performance correlated with the educational content learned.

Platforms for Embodiment and Education

How can embodiment affect education? Embodiment appears to hold much promise for education. The fact that motor codes are still active even after a concept is over-learned, as shown in the Pulvermüller and Fadiga (2010) review, is compelling (and see Hauk (Hauk et al., 2004)). This suggests that if the codes are still sensitive and there are active traces in the sensorimotor cortex, then perhaps we should try to activate those areas again while *teaching* content. We wanted to create a game that would add a motor trace to the act of learning, in this way the game might enhance encoding and retention. A claim could be made that the game should not activate just any motor trace, the game should activate a trace that contains a degree of overlap with the content to be learned. This overlap has been called gestural congruency (Black, Segal, Vitale, & Fadjo, 2012; Segal, Black, & Tversky, 2010). The physical gesture should match the abstract content to be learned.

For the past several years we have been studying which aspects of embodiment account for the most content learned (Johnson-Glenberg, Birchfield, et al., 2014; Lindgren & Johnson-Glenberg, 2013). Embodiment for us means that a gesture has been performed that overlaps physically with the concept to be learned, e.g., the virtual food item is grasped by the learner's hand (with motion capture or a mouse click) and placed in to the avatar's mouth, as opposed to merely hitting an "eat" button (Johnson-Glenberg, Savio-Ramos, & Henry, 2014). We hypothesize that the increased sensorimotor input that occurs during high embodied learning will positively affect learning on both immediate and delayed tests. The study here represents a first exploratory foray into the effect sizes associated with gathering in-game payer data in a highly embodied *Kinect*-based science game.

Other educational researchers have been supporting the use of movement or body-based metaphors in learning before motion capture and games were added to the mix (Nathan et al., 2014). Indeed, Cook and Goldin-Meadow (Cook & Goldin-Meadow, 2006) manipulated children's gesture during instruction on new mathematical concepts. The children who were prompted to gesture while learning retained the knowledge they had gained better than the children who were not prompted. Cook et al. postulate that gesturing serves a causal role in learning, perhaps by giving learners an alternative, embodied way of representing new ideas. Goldin-Meadow states, "…perhaps it is the motor aspects of gesture that are responsible for the cognitive benefits." (Goldin-Meadow, 2011). Nathan and Alibali (Nathan et al., 2014) found a significant relationship between action and cognition and experimenter's language (prompts and hints) as participants learned geometry proofs.

The genesis for our embodied journey. For almost a decade, members of the lab have been designing learning games and simulations for mixed and augmented reality platforms. The term 'mixed reality' was first used by Milgram and Kishino (Milgram & Kishino, 1994). The platform the first three authors have most published on is called *SMALLab* (Situated Multimedia Arts Learning Lab). The *SMALLab* motion-capture platform used 12 ceiling-mounted infrared *Optitrack* cameras to track players' as they moved holding a rigid body trackable object. The experience was very immersive (the projected floor and tracked space was 15 x 15 X 15 feet) and extremely collaborative. Entire classrooms could sit around the projected perimeter and observe and interact with active students in the space. When learning in the mixed reality platform was compared with traditional instruction (teacher and content held constant) significant gains, or trends, were seen that consistently favored the mixed reality platform (Johnson-Glenberg, Birchfield, et al., 2014; Johnson-Glenberg, Birchfield, Megowan-Romanowicz, Tolentino, & Martinez, 2009; Johnson-Glenberg, Birchfield, Megowan-Romanowicz, & Uysal, 2009)

As Connolly et al. (Connolly, Boyle, MacArthur, Hainey, & Boyle, 2012) noted, recent technologies such as mobile devices, online games, virtual worlds and augmented reality on more mobile platforms have had a profound effect on educational gaming. With the advent of more portable and cost-effective sensors, several of us have moved away from the rigid body tracking system, but we opted to continue designing using principles of embodiment. We have created and tested multiple educational games using the Microsoft *Kinect*'s joint tracking algorithm as the input. The *Kinect 1* captures 20 joints on a body at approximately 30 frames per second, we usually have two players up in front of the class moving and learning as the other students observe. Because we are tracking the active students' gestures during play, we can analyze how these gestures, and their veracity, correlate with science-based paper and pencil pretest and posttest scores. This is the driving goal of this article.

Game Design

This article describes two science games, one called the Winching Game and the other called Tour de Force. These were part of a series of games created to help middle school students understand how simple machines - gears and levers - worked. We chose this content because we knew we wanted a series of engaging games for science to keep learners in the science pipeline, we knew it had to meet the U.S. grade standards, and we knew it had to be something that could be intuitively 'embodied'. We reviewed the content of a typical middle school science class and settled on the topic of simple machines. It is attractive because several parts of the body can simulate the action of gears or levers (the arm is a natural lever). After the topic space, was chosen we needed to reframe it through the epistemology of children's scientific thinking. We premised our understanding of a naïve science learner from DiSessa's theory (DiSessa, 1988) that human knowledge consists of many, loosely organized, fragmented pieces of knowledge. These building blocks of understanding are called phenomenological primitives (p-prims) (DiSessa, 1988). They are small knowledge elements whose origins stem from repeated abstractions of very familiar events. We hold that these familiar events are experienced in an embodied manner at the earliest age. For example, many students come to kinematics and other science topics with misconceptions, i.e., that force is not an action with an equal and opposite reaction.

In one of our summer camp pilot studies with 17 middle schoolers, 95% of them incorrectly answered that the largest input gear in a gear train would lift the largest and heaviest item. When queried why, they often replied with something akin to - "bigger is always better" (unpublished data, 2012). With repeated exposure to our games, and scaffolded discovery from the teacher, the goal is to re-organize a student's network of fragmented knowledge elements into a coherent and correct (more expert-like) model. The students may need to "experience" how the construct is incorrect and why. They may need to use *physical* body actions to *virtually* lift smaller and larger objects to really understand why their current gear mental model is incorrect. With this experience, and a new motor memory trace, the student may have a better chance of constructing a knowledge structure that more closely resembles an expert's knowledge structure.

The game design principles. The lab adheres to several design tenets. The design goal is to maximize the following in the games and make them:

- **Embodied** – with as much "gestural congruency" as possible
- **Socio-collaborative-** build for discourse with the playing dyads (pairs of students), also give the observing students tasks to do to keep the whole class engaged
- **Generative** – encourage learners to be active; in real time their choices and gestures affect what is on the screen. That is, they generate what appears on screen/the content
- Give **immediate performance feedback**
- Include **cycle of expertise -** leveling up is classic good game design (J.P. Gee, 2007)
- Include **user-created content –** ideally students should be contributors and not just passive consumers of media. For example, we included an in-game editor that allowed students to create personalized virtual race courses for their peers. This is highly generative and we have seen that it encourages students to take 'ownership' of the content (Johnson-Glenberg, Birchfield, et al., 2014)

The three games described in this article, the code-based drivers, and teacher's guides can be downloaded from www.smallablearning.com under FLOW scenarios; to play a free example of the Ratio Match game on the Kinect V1 go to http://www.embodied-games.com/games/all/view/gear-ratio-game.

The Beauty of Gears

Simple machines are an important application of the middle school science curriculum and content standards. To understand gears, we wanted to encourage students to explore the relationships between a number of "embodiable" concepts including gear size, speed of rotation, and direction. One challenge we confronted was how to embody the idea of diameter (size) and mechanical advantage when two gears interact. Some of the earlier questions on the test were designed to measure students' conceptual coherence and were influenced by an earlier Metz study (Metz, 1985). Her participants worked with a set of physically manipulable gears fashioned with a wooden crank. Two of the gears were marked with the form of a man. When a marked gear was turned clockwise, the man pictured on it somersaulted feet-first. Participants were instructed to arrange articulating gears so that they could "make the two men somersault in the same direction." Via trial and error, participants learned the relationship of parity between gear elements. E.g., if there were an odd number of gear elements between the marked gears, then the two marked gears turned in the same direction. Eleven and 12 years-olds were able to understand this rule. The Metz study demonstrates that students are able to construct complex knowledge through the direct manipulation of physical systems. Our goal was to preserve the powerful learning that can occur via this type of physical embodied experience and integrate it into the affordances of digital media in a game-like manner.

Gears also allowed us to introduce, in a playful manner, the concept of mechanical advantage (MA). A robust energy concept is central to an understanding of all science and MA as well. Although treated as if it is a straightforward and easily defined quantity, energy is notoriously difficult for people to understand. The second author was a physics teacher for 25 years and often her high school and college students confounded energy with the concepts of force and power (and even speed). Energy is usually defined simply as "the ability to do work," a more general and useful definition is "the ability to cause a change." One reason to use the more general definition is that the term work is one that is often misunderstood as well. There are two things that can be done with energy—it can be stored and/or it can be transferred. The transfer of energy by means of work happens by exerting a force on something across some distance. The product of the input force and the distance travelled is the work done on the object. Working is the mechanism for energy transfer. Simple machines (e.g., levers and gears) are devices that make work easier by changing the magnitude and/or direction of a force and the displacement of the object to which the force is applied. This allows for useful work to be done on some object that would be difficult, or impossible, to accomplish otherwise. Examples of tasks that are hard for humans to do alone are lifting heavy boulders, moving pianos up stairs without pulleys or planes, and cycling up very steep hills.

Gears are a good choice for study, besides being able to be embodied, they involve both mathematical and scientific concepts. Gears require the application of both descriptive and causal conceptual models. Their motion and action are familiar (bicycles, pulleys, etc.) and easy to visually perceive because there is no hidden mechanism at work, but gears also reveal stubborn misconceptions regarding both simple (e.g., size, speed, and direction) and complex (e.g., mechanical advantage) constructs. In the games for this series, students were only exposed to double gear systems. There was one input gear and one output gear; players controlled the input gear's diameter. The output gear size was always preset and constant throughout the game. The mechanical advantage construct is a ratio, and it is Output to Input, or O:I. That is, diameter of the output gear divided by the diameter of the input gear.

The lab created a series of six simple machines games, three of these were games on levers. The three lever games were of such short duration (less than one minute) and only resulted in

Figure 1. A dyad in the Tour de Force biking game. The Kinect sensor is in front of the interactive whiteboard. Players straighten their arms and spin the wrist joint in a circle around the shoulder altering size and direction of the input gears, in this case, the pedal gears

two lengths of the arm being captured by the sensor that a meaningful frequency of change in a lever game was difficult to extract. Thus, we focus on only the gears games for this paper. Figure 1 shows the standard stance for the gears games with the arm extended out in front. Turning the whole arm (that is spinning the wrist joint around the shoulder joint) would make the input gear spin in the same direction and would control the input gear's diameter size.

The Hypothesis. Our hypothesis was that the students who were uncertain or confused about the optimal gear diameters for mechanical advantage (MA) would be the ones who switched gear size more often. In addition, those who switched gear size more often during play would be the ones who demonstrated the lowest scores on the traditional paper and pencil knowledge assessments. This means there should be a negative correlation between number of gear switches and scores on the pretest, demonstrating that those who started with lower prior knowledge did not perform as well during gameplay. In addition, there might also be a negative correlation on posttest, because poor in-game performance should translate to poor posttest performance. The theory is that if a student really understands the tenets of MA and optimal input gear size for a task then s/he will go straight to that gear size and not use trial and error (bounce around to many diameters). Thus, a student with fewer gear switches (lower frequency) during play would be a student who understood MA and show greater gains on the gears knowledge tests. This is a timely question because if we can make "construct sensitive" games and gather meaningful in-game data to predict the amount of content learned, then we do not have to waste learners' class time giving repetitive or summative paper/pencil tests.

METHODS

Participants

Participants were 23 7th graders from a private middle school In America with 160 students. The study began with 25 participants, but two (one from each class) were absent the day of the posttest, or their in-process data were corrupted. Fifty percent at the school received financial aid and 52% of the students were described as "people of color" by the Principal. There were 11 females in the two classes. There were no significant knowledge test differences due to gender.

Two science classrooms were used in the study and the same teacher covered the same content in both classes.

Procedure

The entire study was a seven day-long intervention on Simple Machines. The study began with the concept of gears and then moved onto levers (this article does not cover levers). There were three days on gears. On Day 1 a pretest was given and then students played the Ratio Match gears tutorial game. On Day 2 students played the Tour de Force game, on day 3 they played the more difficult Winching game. On day 4 they took the posttest and then continued on with the levers content. Participants were quasi-randomly assigned to dyad (more on this in the Discussion section).

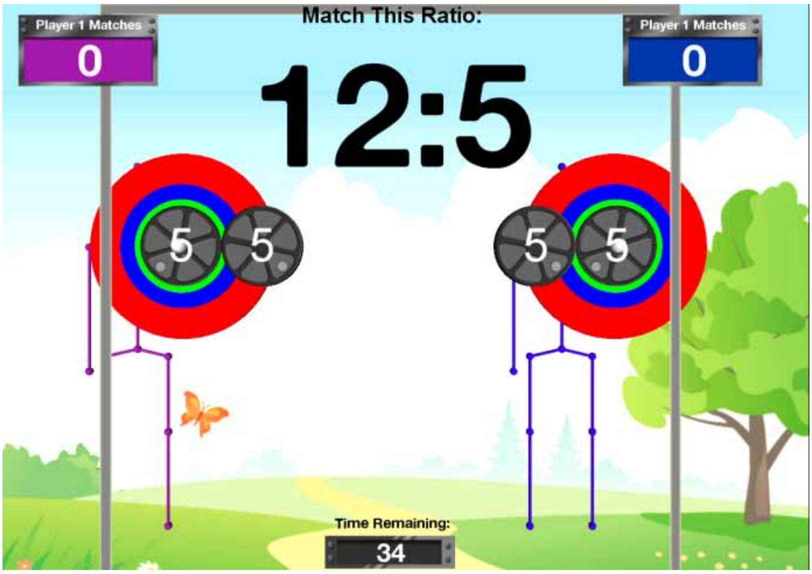
The classroom lessons were co-designed with two science teachers, the lead programmer, a cognitive psychologist, an experienced game designer, and a physics subject matter expert. The classroom teacher in this study was engaged and very innovative, e.g., he brought in his own manipulatives like planks and bricks for levers, and his own bicycle to demonstrate gears. Thus, the content for the lessons was not solely based on the *Kinect* games we supplied, those games were supplemented with several short lectures given by the teacher. The teacher was provided with a scripted guide. (See example of guide for the biking Tour de Force game at http://smallablearning.com/wp-content/uploads/2014/08/le_tour_de_force_teacher_guide_flow.pdf.) The teacher occasionally departed from the guide and we allowed him this leeway to keep the learning authentic, he did the same for both classes. He did not veer from the *Kinect* game scripts we supplied when using the games.

Day 1- Gears Tutorial Game. After the pretest, student dyads took turns learning the mechanics of the *Kinect* using the Gear Ratio Tutorial Game. Two players either volunteered or were asked to come to the front of the room. The regular classroom projector projected the image (approximately 80 inches diagonal) on the wall. The two players, the dyad, would stand in front of the *Kinect* sensor with their backs to the class and practice matching gear diameters in this sandbox-style tutorial game. (Video available https://www.youtube.com/watch?v=kSsiJZOUKt4&feature=youtu.be.)

Figure 2 shows the screen shot of the tutorial *Ratio Match* game. The *Kinect* is tracking two key joints on each players' bodies, i.e., the wrist and shoulder. The shoulder is the pivot point. Spinning of the arm results in the the wrist joint rotating around the shoulder joint with a certain diameter.

The input gear rotates in real time with the player, or at same speed on screen, and in the same direction as the player's arm. This is a strong example of gestural congruency. The virtual input gear diameter snaps to three different sizes. Thus, if a player makes a very wide circle around the shoulder joint it will result in the largest gear diameter of 12- the outer, red area. In Figure 2, the players are being prompted to match to a ratio, in this case an input size of 12, and the output gear on the inside is locked at a diameter of 5. Thus, 12:5 is the target ratio. The player that first completes two revolutions with an input maintained at 12 will win the point. The game ends when the 45 second timer runs out. The total number of matches are presented in the boxes on top. (Note: we did not gather in-process data on this short tutorial.) Although the *Kinect* sensor can track up to four unique players, we never design for more than two players since our games often use gross physical movements and students need to be far enough apart from each other so they do not hit one another by mistake. Everyone in the class got to play each game more than once. This is a sandboxy, "no-stakes" game to get the players used to the *Kinect* game mechanics.

Figure 2. Screen capture of the Ratio Match tutorial game. The gears with colored concentric circles on the edges represent the three variable input gears sizes for each player



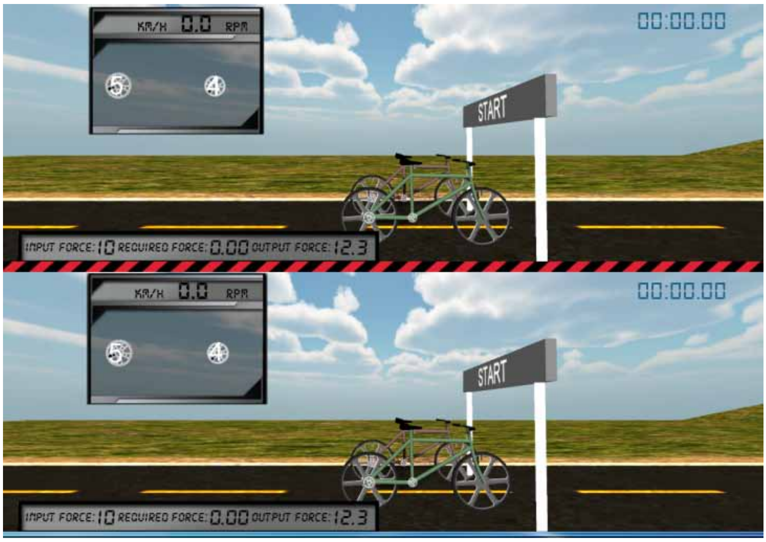
Day 2 - Tour de Force Game. Dyads played the *Tour de Force* biking game on the second day. Again, the teacher first asked for volunteers and then the non-volunteering students were placed together in dyads. Thus, dyads were "quasi-randomly" assigned; they were not preset. The game had a cut-off play time of 240 seconds. On average the last player usually crossed the finish line by two minutes (120 seconds). Again, Figure 1 showed a player with her arm out in front as she turned the input gear on the pink bike on the top of the screen, the bike is going up a small hill. The second player is lagging behind on the green bike.

As players came to the front of the class, the experimenter entered their subject IDs into the computer from a prepopulated list. The size of the input gear, as measured via distance from the wrist to the shoulder pivot point, could vary (snap to) three sizes. The diameter changes were time stamped and mapped to subject ID. Figure 3 shows a close up of the opening screenshot.

Players are also able to see statistics on their performance such as input force, output force, kilometers per hour, and time played so far. Figure 3 is a screen shot of the Tour de Force biking game, first one to cover the hilly course and cross the finish line wins.

At the beginning of the lesson the teacher guided student exploration of the equation: *Work = Force X Distance*. The lecture included the concept that there was a limited amount of input force, that is, they had only so much effort they could exert on the input gear. There were three input gear sizes to choose from 4, 5, and 16.The winner was the one who crossed the finish line first. To be the fastest, a player needed to show facility with critical constructs associated with the input gear (the pedal gear). First, on the flat section of the race course the largest gear (size 16) should be used. Second, on the steep hill sections, the smallest gear (size 4) should be switched to for optimal performance. The gear switch needs to be well-timed and maintained as long as the player was on that terrain. If the player was not generating enough output force the bike will remain stationary. It did not slide backwards.

Figure 3. Screen shot of the Tour de Force biking game



A learner who had truly intuited mechanical advantage would exhibit both a very fast time and a minimal amount of gear switching. We created a course with 16 waypoints, that is, places where the hill slope could change. All participants played on the same default course with the same hills. On the far right in Figure 3, you can see the beginning of a hill.

Configuration Panel - Tour de Force. Because we strive for players to also be creators, or generative in our games, we included a configuration panel, accessed with the key strokes "ctrl c". The panel (see Figure 4) allowed the students - or teacher - to alter the slopes of the hills by varying the Y coordinates. In this way slopes that were impassable could be created and discussion could occur around math concepts like graphing, mechanical advantage, and game design, e.g., what makes for a "fun" and challenging game, versus a simply frustrating experience (not being able to ascend a hill of 80°). Pedal force (input force) and bike weight could also be altered in a different section of the configuration panel.

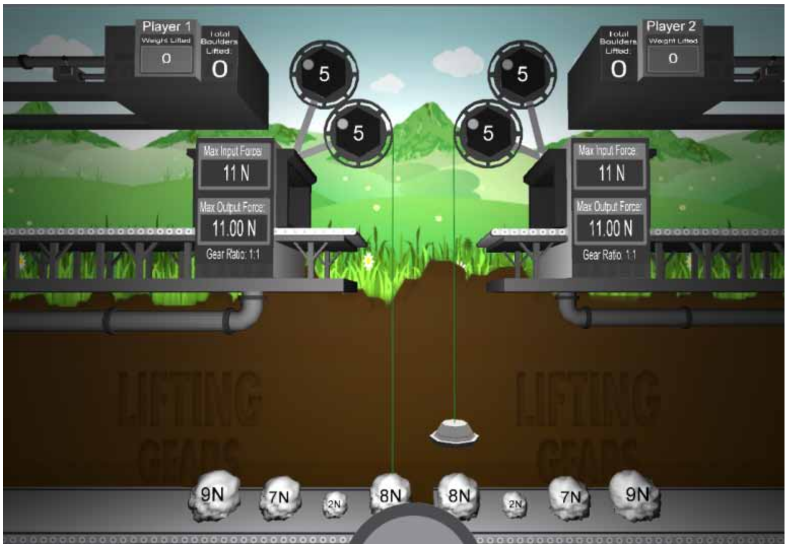
Profile of the Player. Players come to games with varying amounts of prior knowledge, both in terms of comfort with game play and knowledge of the content to be learned. Pulling these issues apart can be difficult with traditional games, especially if multi-button traditional game remotes are used. Mislevy, et al. (2012) make intriguing points about how game designers and statisticians can work together in designing games to mitigate some of the prior gameplay knowledge that could add noise to a participant's knowledge gains. They stress how the assessment design framework called Evidence Centered Design (ECD) can complement game design principles, so that designers can address assessment criteria such as reliability and validity jointly with game criteria like engagement and interactivity in mind (Mislevy, Behrens, DiCerbo, & Levy, 2012).

Regarding prior gameplay knowledge, we posit that the game play mechanism in this study was novel for all participants since the *Microsoft* Kinect for *XBOX* suite at the time did not have any products similar to our gears games (and as of 2015 still does not). It was novel, yet intuitive; players did not need to use cognitive resources to recall the difference between buttons A and B, merely that a wider arm circle represented a wider gear diameter.

Figure 4. Example of a configuration panel for the Tour de Force game, hill slopes can be changed by the players

Regarding prior content knowledge, middle schoolers often approach gear trains with a knowledge structure that includes a misconception regarding diameter and which size input gear would be most efficacious given the circumstances. Previous research supports that students have little understanding of mechanical advantage in gear trains. One way we measured whether they understood that they needed a smaller input gear to get up a steeper hill was via the number of diameter switches they made during play. Students often attempted many different gear diameters before they began to address their most common misconception - "bigger is always better". The students who either understood mechanical advantage on the bike gear train, or who could use the formula of the ratio of O:I (output to input) made the switch quickly to the smallest input

Figure 5. Screenshot of the Winching Game. The input gear that can be altered by three diameter sizes is the one on top; the bottom, or the output gear, is locked at 5



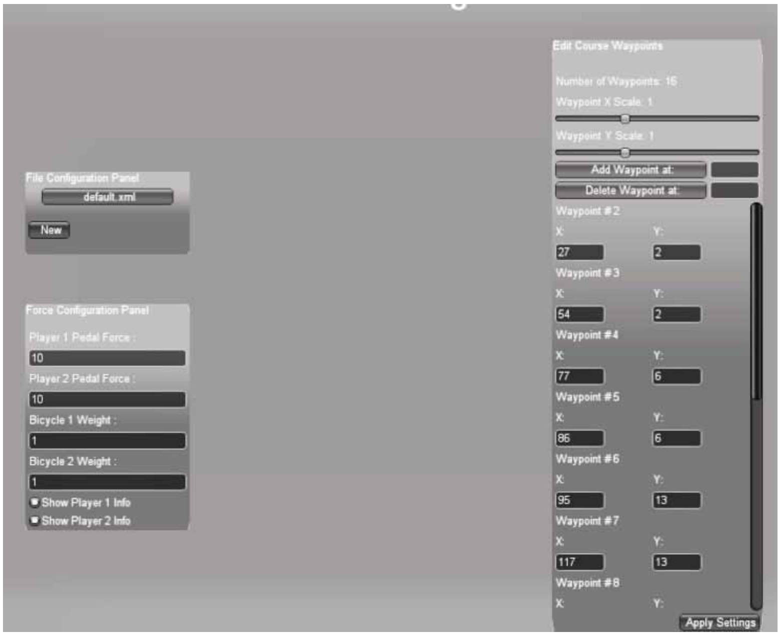
gear size and stuck with it until they had reached the top of the hill. At the top, when the course usually flattened, players who got it would switch adroitly back to the bigger gear and maintain that diameter appropriately.

Students who did not understand the concept would bounce around trying all different sizes of input gear until the bike finally moved. We labeled them the "Bouncers". The students who were more efficient, and consistently demonstrated they got it - were labeled the "Got its". It appeared that all students attacked the first couple of hills with the type of exploration technique commonly seen in mastering new games, but the Bouncers never really moved from exploration to exploitation during these games. We wanted to know if some of this behavior could be captured via in-process data and if overall number of switches would correlate with other subject variables like prior knowledge.

Day 3. The Winching Game - On the final gears day, dyads played the Winching Game. Please see Figure 5. The goal was to lift boulders of varying weight (force) from a pit onto a conveyor belt. The player with the most boulders lifted when the timer buzzed at 90 seconds won. In Figure 5 the input gear is the one on top. The output gear was locked at 5 again. When the students played the Winching Game they needed to realize they should change the size of the input gear to winch up the larger boulders efficiently. The input gear sizes were set to 5, 7 or 9.

One aspect that made this game harder than the previous biking one was that the input gear now moved in two directions. Because time was of the essence, the optimal strategy would be to use the largest gear to lower the winch because it moved the magnetic head down the fastest (spin to right). Then, when the magnetic head attached to the boulder, the student must deduce which size input gear would be most efficient to lift the boulder upwards. After that, the student must spin in the opposite direction (spin to left). Boulders were randomly seeded at the bottom of the screen and moved on a second conveyor belt, they ranged in weight from 2 Newtons (N) to 9 N for this game. Newton is a measure of force and for this game can be thought of as

Figure 6. Example of a configuration panel for the Winching game



equaling .22 lb. or .10 kg. One of the constraints was that the largest input gear would *not* lift the largest boulder. The boulder would be stuck no matter how rapidly the player spun his or her arm. Again we observed that the Winching Game was more difficult to master than the Tour de Force game because the input gear needed to spin in two directions not just one. (See video on www.embodied-games.com website or http://www.youtube.com/watch?v=NHLwQ8kZQ5A .)

Configuration Panel - Winching Game - This game also included a configuration panel so that students - or the teacher - could alter start conditions and discuss how that affected play, mechanical advantage, etc. Figure 6 shows the configuration panel for the winching game. Players could control amount of lift force per player, amount of play time, range of boulder weights and how much data appeared on the screen.

Measure

Gears Knowledge Measure. This measure was an experimenter-designed test created with several middle school teachers and a physics subject matter expert. It was pilot-tested several times to be age-appropriate. There were 13 items to be answered on the gears test, either multiple choice, open-ended, or fill in the blank items. The very last item queried students to choose the correct relatively-sized gear to winch up objects of varying mass. The maximum score achievable on this subtest was 54 points. Multiple choice questions were worth three points and open-ended questions ranged from zero to five points. It included both near and far transfer items. The invariant test is in the Appendix. The highest score achieved on the posttest was 37.

The Gears Lessons: Learning Goals

We were explicit about several instructional goals in the lesson. The full teacher's guides can be found at the previously mentioned websites.

1. Understand energy interactions in terms of transfer and storage
2. Develop a concept of work as a mechanism for energy transfer
3. Apply the concept of conservation of energy
4. Understand and demonstrate comprehension of calculations of the efficiency of a simple machine, specifically mechanical advantage via game play
5. Define and show comprehension of mechanical advantage (*the factor by which the input force is multiplied by the use of a machine to transfer energy*)

RESULTS

Knowledge test. All results were run with SPSS v22. All alphas levels have been set to .05 and were two-tailed.

Pretest and Classroom Differences. The two classes differed significantly on gears knowledge pretest. The teacher was surprised by this result because he had been teaching the students for four months and stated the two class skill-levels were equally matched. The only difference he suggested was the before and after lunch distinction. The AM class scored 20.08 (SD = 6.56), and the PM class scored 12.08 (SD = 9.89) on the pretest. An independent *t* test revealed a significant difference, $t_{(22)} = 2.32$ $p = .03$. Interestingly, the scores by posttest were no longer significantly different by class, $t_{(22)} = 1.57$ $p < .14$, and a General Linear Model analysis using the interaction of time by class was not significant, $F_{(1,20)} = .08$. Thus, the decision was made to combine the classes into one dataset to increase overall power for the study since all students received the exact same instruction by the same teacher.

Posttest. As a larger group, the gears knowledge test results for the 23 participants were: pretest $M = 16.81$ ($SD = 8.89$); posttest $M = 21.19$ ($SD = 7.90$). This gain was significant on a paired *t* test: $t_{(22)} = 2.33$, $p = .03$ (also confirmed in a GLM analysis run with class and time, the time variable was significant, $F_{(1, 20)} = 5.18$, $p = .03$. The effect size was medium, Cohen's *d* is typically reported in the cognitive sciences. It is the mean difference between scores divided by the pooled average of the pretest and posttest standard deviations, $d = .52$.

In-Process Analyses – Tour de Force – Entire Sample. Our hypothesis was that the students who were more efficient with their gear changes, that is, the "Got Its" with fewer switches during game play, would do better on the paper and pencil test. The metric used was gear diameter switches - the average frequency of change of gear size during play. The game cut off after four minutes of play, during the first play students interacted with the *Tour de Force* game for a mean of 78.41 seconds (Range = 32 to 190, $SD = 36.37$ seconds). During this time students averaged 72.27 gear switches (SD = 41.52 switches). We predicted that the "Got Its" who understood mechanical advantage would use fewer switches and would cross the finish line faster and this was evident with the entire sample. Overall those with shorter times used fewer switches, $r(22) = .82$ $p < .001$. But, were these faster players the winners within their dyads?

Tour de Force - Dyadic Analysis. Tour de Force lent itself to analyses using the pair of players (the dyad) as the unit of analysis because the game course did not alter between dyads. We saw that if the winner was faster, the loser was comparatively faster as well within the whole sample. Thus, partnerships affected play across the sample. However, we only had 11 dyads

Table 1. Paired and Independent t Tests Comparing Dyadic Winners and Losers

Dyad Comparison	df	M difference (SD)	Cohen's d	t test sig. level
# of Gear Switches (Loser - Winner)	10	39.65 (41.91)	.95	Paired = 3.14** *p* = .013**
				Independent = 2.03 *p* = .032*
Knowledge Test Gains Score (Loser Gain – Winner Gain)	8	- 5.78 (10.40)	.56	Paired =- 1.67 *p* = .13

on the day Tour de Force was played. (We only used first time play data and one student went twice, with a first time player and so we did not include that dyad.) We note that statistical power is an issue in the following analyses, indeed, *G Power* version 3.1 provides an estimate of only .26 to find significance at a .05 alpha level using a .50 effect size; if we wanted .80 power we would need 34 pairs.

With more pairs, we would be able to consider using other types of analyses like Hierarchical Linear Modeling (HLM). Although the case could be made that these data are not traditionally nested. HLM assumes the structure of the data are nested at a level we never considered in this small study, e.g., teacher, school, or district. It may be the case that position of play (first or last pair to play after observing) had an effect on scores and time, but we did not have those data. We do consider the paired to be linked, or yoked, within their game space. In this manner each dyad player is not independent (i.e., one wins – ergo the other *must* lose). An analysis of interest might be whether the number of gear switches separating the winner versus loser within dyad was significantly different, and that analysis was run.

Analyzing the set of first-time play dyads, the winner switched frequency of gear diameters on average 51.91 times (SD= 20.67), the loser switched 91.55 times (SD= 53.13). When students were analyzed as the winner (coded as 1) and loser (coded as 0) within dyads, the frequency of switch between players was correlated as well, r = .68, p =.02. Faster dyads had fewer switches overall and some of the losers in a faster dyad would have been the winners in a slower dyad. Again, partners affected each other's' play. We wanted to know if the difference between switches was statistically significant regardless of whether the yoked pair were in a relatively faster or slower dyad. Using a *paired* t-test within dyad we account for some variance between dyads. But because this may seem controversial, we also ran independent t tests. The paired t test revealed that the difference between the winners' and losers' switch frequency was significant, see Table 1, p = .01, with a large effect size (Cohen's d) of .95. Independent and bootstrap analyses are also reported and were also significant supporting the hypothesis that winners used significantly fewer switches within play, Bootstrap t = 2.03, bias = -.02, p = .03 (SE = 11.70).

In Gameplay and the Knowledge Test. The next question was whether the winners in the game showed greater gains on the knowledge posttest. The effect of change on test scores was moderate, d = .56. See Table 1. The inferential paired t test was not statistically significant because we only had 8 dyads in that analysis (several players missing data or posttest knowledge scores). We note again the power issue, but the direction of the t value would support the hypothesis that winners in the game dyad did better on the knowledge posttest than the losers within a dyad. There was a sort of "switchover" seen as well. That is, the dyad losers did better on the pretest (by two points on average) but *worse* on the posttest (again by two points on average). This may suggest that in-game processes were capturing knowledge as it changed. Winner and loser knowledge pretests were not significantly different within dyad (paired *t* value < .80), perhaps suggesting that friends who score similarly on tests often choose each other to play with, nor was amount of switching significantly correlated with pretest scores within dyads, $r < .25$.

Regarding knowledge gains an independent *t* test on the gains between losers and winners was also run at a reviewer's request and the *t* value decreased to negative .95 from negative1.67.

Tour de Force Gameplay and Knowledge Test - Whole sample. Using the sample as a whole group we predicted a negative correlation such that the students who made fewer gear switches would have higher test scores. There was significant evidence of this on the pretest, $r = -.41$, $p = .05$; however, at posttest this negative correlation did not hold, $r = .12$, $p = .29$, NS.

2nd Game - In Process Analyses – Winching Game – Whole sample.

In the Winching Game students needed to spin their arms in two directions, this game was reported by players to be somewhat more difficult than the first game. During the Winching Game the students lifted on average 17.7 rocks ($SD = 8.47$). All teams played for the same amount of time, the teacher never altered the game play time (from the configurable panel). The range of gear size switches was 43 – 140; $M = 75.28$ ($SD = 27.61$). Gear switches was not predictive of number of rocks successfully lifted, $r = -.10$. Although, in this small sample we still see that the valence is negative.

Gameplay and Knowledge Test - Winching Game - Whole Sample. We predicted a negative correlation between number of switches and test scores, such that the students who made fewer switches, would have higher test scores. We saw some evidence of this on the pretest, $r = -.26$ (NS) and stronger evidence on the posttest, $r = -.37$, $p = .07$, which represents a statistical trend.

We did not run the same of sort of comparative dyadic analysis on this game because each player within each dyad was confronted with a different game course. That is, rock sizes were randomly seeded for each game and for each player on the bottom conveyor belt. It was possible for three rocks of the SAME size to come out in a row making one player's task much easier than the other player's task (i.e., the second player may have been faced with varying rocks or a streak of large N rocks). This is different than the Tour de Force Game where the hill series never varied within or between dyads.

DISCUSSION

Both the gears Winching Game and Tour de Force Game were designed as playful environments for students to practice and demonstrate their understanding of gear trains and mechanical advantage. On average the two classes demonstrated statistically significant gains in learning on the knowledge posttest. The games were embodied and innovative in that they used the *Kinect* sensor as the input device so the body could mimic the tool of instruction (gear or lever). Our goal was to make the learners' movements map to the content to be learned with gestural congruency. We mined student performance during gameplay to explore how "physical" arm rotations and "virtual" gear diameter shifts related to dyadic performance and on more traditional paper and pencil tests. The faster students were always the winners in the dyad and the faster students also used significantly fewer gear switches (at least in nine of the 11 Tour de Force dyads). A key research question is whether in-process gameplay correlates with performance on more traditional tests. The results suggest that this may depend on the difficulty of the game. In the easier Tour de Force game where arm spin rotation was in one direction, the "Got Its" did significantly better on the pretest. However, players' in-game performance was not predictive of posttest, that is, students' knowledge post-intervention was not correlated with diameter gesture-choices during learning. For the more difficult Winching game the valences of the pretest and posttest correlations remained negative, as we had predicted. By posttest the better learners were generally using fewer switches representing a trend for more learning ($p < .07$). Thus, there is some evidence that in-game performance on appropriately calibrated (i.e., harder, more effortful) games can reveal a learner's profile between two time points.

These findings suggest that when students are in a more challenging game, one that might match their Zone of Proximal Development (Vygotsky, 1978), they might exhibit patterns of movement that suggest ongoing comprehension. The negative correlations support the hypothesis that the "Got It" players, the ones who rapidly understood how mechanical advantage worked during the game, also showed greater gains on the posttest. These were the ones who could see a hill coming and switch expeditiously to a smaller input gear to get up the hill, and then switch back to and maintain a larger gear when on a flatter slope.

Within a dyad, the winners in the dyads did not start with statistically higher pretest scores (as a simplistic prior knowledge-based hypothesis might predict). However, when the winners took the posttest they did on average have higher *posttest scores* than the dyadic losers. This suggests that the game may have been capturing some moments where knowledge may have "switched over", that is, a time when pretest was no longer the sole predictor of performance – at some point during the three day intervention a *different sort of learning* may have been occurring that was not contingent on the previous knowledge the student arrived to the task with. Or, it may be that the quasi-random nature of the dyadic pairings added too much noise, e.g., "smarter kids hang together" so there was self-selection bias and these were the first to volunteer, so that when whole group correlations are run many nuances are lost. A final, equally speculative, assumption might be that the paper and pencil static test may be capturing a different sort of declarative, or crystalized, knowledge that varies from what gestures can show. Gestures can reveal knowledge that is internalized, but not captured on other types of tests or present in speech acts (Goldin-Meadow, 2014). We have recently begun to gather movement data using a touch-sensitive assessment tool, a large WACOM Pro tablet (Johnson-Glenberg, in preparation).

Larger studies are needed, as well as a low-embodied control group. Nonetheless, this study is a beginning in gathering the sorts of effect sizes and results that can be seen in embodied science games. Learning scientists have speculated for years regarding the constraints associated with gathering validity and reliability on gameplay data. One timely question, given a test-besotted school environment, would be: "Is it possible to gather valid knowledge information about comprehension in-vivo during gesture-based gameplay?" Our results suggest "*yes",* in ludo data can be predictive of knowledge. If this is the case, then why should we force students to work through lengthy paper and pencil tests post-gameplay? If a student demonstrates in real time that s/he has mastered the concept, then a traditional multiple-choice type test need not be administered (J. P. Gee & Shaffer, 2010). We view these sorts of games-as-assessment as a highly efficient use of classroom time.

Limitations and Future Directions

The *Kinect* sensor was used as the input device; however, these sorts of movement and data-driven decision analyses can be accomplished with other technologies like the *Leap Motion*, the new *Intel RealSense* embedded camera, or traditional mouse/touch screen input streams as well. One reason the lab designs with the *Kinect* is that we want to elicit larger amounts of sensorimotor activity from the user. The theory (Johnson-Glenberg, Birchfield, et al., 2014) is that the greater the amount of efferent motor activity, then the stronger the memory trace will be (assuming the gestures meaningfully map to the content to be learned). Thus, we have created games that use the whole arm to spin a gear, but circling the finger to spin a gear on a tablet may be just as effective. The amount of sensorimotor activation is a question for future studies, as well as the degree of gestural congruency. We agree with Nathan et al.'s (2014) observation in a recent paper that also assessed gear knowledge. The authors state that "...grounding actions may be most effective when the underlying mathematical ideas...align with the physical and spatial relation..." (p.192). Nonetheless, the gesture they used of "tapping on" virtual gears to learn the parity rule, may not have been as powerful as using a full-arm gesture to simulate turning the virtual gears in real time, as we did in our study.

New assessment metrics. This emerging field of game analyses is in need of more adroit, more sophisticated inferential tools. We were uncertain how to deal with the dynamic fluxes occurring over the class-long hour of play. For example, it appeared that many of the beginning-of-play dyads were not truly randomly assigned (e.g., friends), but the end-of-play dyads were randomly assigned by the teacher from a pool without replacement. Perhaps we could add decay function to an analysis to account for increasing randomness or independence of the pairs? On a separate note, the end-of-play dyads did observe more bouts of play. In this manner, one might expect the end-of-play dyads to pick up faster on the optimal mechanical advantage gear switch. Anecdotally, this was not seen in either the gears or levers games. The end-of-play dyads were generally composed of the slower and poorer performing players. One hypothesis is that they are the shyer, more uncertain, students and so we recommend a random number generated sheet be used by the teacher in these sorts of situations.

Clearly this exploratory study needs more dyads for power considerations. In addition, with a longer time series not binned by events Hurst exponents and persistence measures could have been reliably gathered. Hurst exponents are interesting measures for predictivity and time series. As an example, in a social experiment with children, DiDonato (DiDonato et al., 2012) showed that when young children demonstrated flexibility and nonpersistence (as in, a Hurst exponent closer to .5 which connotes more randomness and less pink noise), the exponent was a positive indicator for later behaviors. They found that preschool children who were more gender flexible with play partners during earlier play showed better positive adjustment on several scales six months later. In the learning and computers literature, Snow et al. (Snow, Allen, Jacovina, & McNamara, 2015) investigated how log data can be used as a proxy for self-regulated learning and agentic behavior. Specifically, they identified patterns of behaviors that indicated controlled and ordered processes as students made choices through two computer-based tutors using various dynamic time series methodologies (i.e., Hurst, Entropy, Random Walks).

The Hurst exponent from a time series is certainly a more nuanced metric than the measures of central tendency used in this paper (Means, etc.). With a longer series we would have been able to use a delta of variance as a meaningful change over time metric. What might that look like? We will borrow terminology from Stafford and Dewar (Stafford & Dewar, 2013) who categorized players as either *Explorers* (what we might call Bouncers) or *Exploiters* (what we call the Got Its). Stafford and Dewar gathered statistics on hundreds of thousands of on-line players on a simple perceptual/motor game called *Axon*. They first binned players into percentile groups based on variance during first five times of play (higher variance = *explorers*) and then correlated that with subsequent performance data (plays six through 10: $r = .59$, $p < .0001$). The Explorers were not overly concerned with getting it right the first few times they played. Explorers moved all over the screen and spaced their practice sessions out. The Exploiters massed their practice sessions and were primarily goal-driven. Bootstrapped confidence intervals for the correlation were at the 95% level demonstrating that the Explorers did better at later gameplay even though all in the sample practiced the same amount of time. This is an in-game metric that tells us something about the player profile over time. In the Tour de Force game, performance through gameplay like latency-to-switch-for-hill could reveal intriguing dynamic player profiles over time and further address KSAOs.

Prior Knowledge. We are interested in the construct of prior knowledge and how it interacts with intervention. Low and moderate pre-intervention knowledge students may be able to learn more in an embodied game because the learning is not driven primarily by language or memorizing symbols. These students may demonstrate higher gains when in an embodied condition. On the other hand, it may be the case that using the body is overwhelming at first, sensory overload may occur. Low prior knowledge learners might actually benefit from a symbolic tutorial before attempting an embodied session? In addition, an order of intervention effect may be seen that is dependent on a prior knowledge profile. This point deserves further study. We did not have the power to run such aptitude by treatment interactions (Cronbach & Snow, 1977).

Transfer. Transfer remains a thorny issue. In a relevant gears example, Dixon and Dohn (Dixon & Dohn, 2003) directly instructed participants to use the alternation (parity) strategy on a structurally analogous task in which balance beams were connected end-to-end in a series. Participants were asked to predict whether the last balance beam in the series would go up or down, given the movement of first beam. After being instructed on alternation and solving 10 problems using alternation, participants were given gear-train system problems. Despite immediate prior mastering of the alternation strategy, participants showed no evidence of transfer. Their median discovery trial did not differ from that of uninstructed participants. In our study, the final test question (item 10) asked about relative input gear size when winching up objects of varying masses. These were items to be moved into an apartment, and not boulders like in the Winching Game. On the pretest only three participants chose the correct answer for the heaviest object (the mattress, 13%), on the posttest only six chose the correct answer (25%, paired t <.80, NS). This was not the sort of transfer gain we had hoped to see. It reminds us that teachers may still need to be very explicit with students about what was learned, and perhaps remind students to transfer and apply the MA knowledge to other similar content (unless the content is overlearned).

Adaptivity. For future design, we will work towards tracking behavior and integrating what we know about in ludo performance and placing learners in appropriate levels. If the student is a Bouncer and remains a Bouncer throughout several games, then the system should be able to place the student into a tutorial that goes over the concept of MA again, or flag the teacher to come and make sure the student fully understands the mechanics of the physical gameplay and/or the concepts.

Observation as an IV. Many of our games are designed with a performative aspect. The student(s) go to the front of the class to play and demonstrate their knowledge. There may be effects related to position or time of play. As in, the final students have observed more play and seen the mistakes of the earlier players so they should be at some advantage. We did not have position of play readily available to analyze. Previous ranked player analyses have not shown a significant difference due to time of play (Johnson-Glenberg & Hekler, 2013), although that was also a small n study. We do know that after 20 minutes of observation the students who have already played these short games begin to get restless. We now recommend that teachers use these short games as "stations", and not make the entire class watch for more than six or so sessions.

Students as Creators. Finally, we are excited about building in-game editors and allowing students to alter gameplay for peers. We would like to understand why the configuration panels were not fully used. For the final few instructional minutes of the Tour de Force game, the teacher demonstrated to the students the panel. He changed the Y axis on one hill so that the extremely steep hill would be impossible to ascend. However, the teacher did not let the students explore this on their own; he did not encourage them to build or sketch out separate race courses for different teams to play. One idea for the next classroom instantiation is to add code so that the game will not advance after five play sessions unless some of the hill parameters are altered and played through. Teachers often report feeling a press to get through topics and so changing around the game and going deeper to explore different start states may seem like a luxury to them. A hypothesis worth testing is whether students in the "creator condition" retain more information than those who never alter the game or create varying start conditions.

CONCLUSION

Two games called Tour de Force and the Winching Game were designed to instruct middle schoolers in the concepts associated with gear trains. Learners used the body to map the relatively abstract concept of mechanical advantage to physical, kinesthetic sensations. The games used the *Kinect* sensor as the input device to track the changing diameter of the player-created input gear. Paper and pencil tests were administered before and after the game intervention and significant gains in learning were made. In addition, dyadic data were gathered during play regarding amount of gear switches made during play. Data were systematically examined to understand student movement performance and explore how the arm rotations and gear shifts related to scores on more traditional paper and pencil tests. Negative correlations were predicted, such that, players with fewer gear changes would score higher on the tests. The valence and magnitude of the correlations between gear switches varied between the two games. For the easier first game, movement data significantly negatively predicted pretest score, but not posttest score. For the more difficult second game, The Winching Game, gear switches were negatively correlated with both pre- and posttests.

These exploratory data provide a window onto how students might perform on traditional tests. One take-away for our lab was that not all embodied games are created equal, even though the same team created the games and play-tested them with individual students. The "in the wild" classroom students reported that the Winching Game was more difficult to master. It may be that the predictive effects of games emerge only from a game that is sufficiently challenging, or in the learner's appropriate Zone of Proximal Development (ZPD) (Vygotsky, 1978). If a game is too easy, it has low sensitivity, then being a winner or loser in a dyad will not reveal much about differences in comprehension. We also make note that the game might be capturing a different sort of knowledge than the crystalized paper and pencil test. Goldin-Meadow contends learners' gestures "precede, and predict" the acquisition of structures in speech (Goldin-Meadow, 2014). Thus, the learner may be understanding the concept and gesturing adroitly in the game, but still not be able to demonstrate that comprehension on a symbolically-oriented assessment measure.

Although the study sample size was small for inferential statistics, effect sizes that might be associated with short in-class, embodied games are reported. One primary goal is to use immediate game-style feedback to attenuate the need to give repetitive, time-wasting paper and pencil tests. We find this to be a promising intersection of gesture-based STEM instruction and in ludo assessments.

ACKNOWLEDGMENT

Funded by the Gates Foundation/NGLC Wave II Grant. Special thanks to Daniel Figueroa, Cyndi Boyd, and Masumi Iida.

Correspondence concerning this article should be addressed to Mina C. Johnson-Glenberg, at Behavioral Sciences Institute, Radboud University of Nijmegen, the Netherlands, 6543JA. E-mail: m.johnson@pwo.ru.nl

REFERENCES

Barsalou, L. W. (2008). Grounded Cognition. *Annual Review of Psychology*, *59*(1), 617–645. doi:10.1146/annurev.psych.59.103006.093639 PMID:17705682

Black, J. B., Segal, A., Vitale, J., & Fadjo, C. L. (2012). Embodied cognition and enhancing learning and motivation. In D. Jonassen & S. Land (Eds.), *Theoretical foundations of learning environments*. NY: Routledge.

Clark, D., Brady, C., Sengupta, P., Martinez-Garza, M., Adams, D., Killingsworth, S., & Van Eaton, G. (2014). *Evolving and balancing formal and informal representations.* Paper presented at the Proceedings of the International Conference on the Learning Sciences, Boulder, CO.

Coller, B., & Scott, A. A. (2009). Effectiveness of using a video game to teach a course in mechanical engineering. *Computers & Education*, *53*(3), 900–912. doi:10.1016/j.compedu.2009.05.012

Connolly, T. M., Boyle, E. A., MacArthur, E., Hainey, T., & Boyle, J. M. (2012). A systematic literature review of empirical evidence on computer games and serious games. *Computers & Education*, *59*(2), 661–686. doi:10.1016/j.compedu.2012.03.004

Cook, S., & Goldin-Meadow, S. (2006). The Role of Gesture in Learning: Do Children Use Their Hands to Change Their Minds. *Journal of Cognition and Development*, *7*(2), 211–232. doi:10.1207/s15327647jcd0702_4

Cronbach, L., & Snow, R. (1977). Aptitudes and Instructional Methods: A Handbook for Research on Interactions. New York: Irvington.

DiDonato, M. D., Martin, C. L., Hessler, E. E., Amazeen, P. E., Hanish, L. D., & Fabes, R. A. (2012). Gender consistency and flexibility: Using dynamics to understand the relation between gender and adjustment. *Nonlinear Dynamics Psychology and Life Sciences*, *16*(2), 159–184. PMID:22452931

DiSessa, A. (1988). Knowledge in pieces. In G. Forman & P. B. Pufall (Eds.), *Constructivism in the Computer Age*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Dixon, J. A., & Dohn, M. C. (2003). Redescription disembeds relations: Evidence from relational transfer and use in problem solving. *Memory & Cognition*, *31*(7), 1082–1093. doi:10.3758/BF03196129 PMID:14704023

Gee, J. P. (2007). *Good Video Games and Good Learning: Collected Essays on Video Games, Learning and Literacy*. New York: Peter Lang Publishing.

Gee, J. P., & Shaffer, D. W. (2010). Looking where the light is bad: Video games and the future of assessment. [Phi Delta Kappa International]. *EDge*, *6*(1), 2–19.

Glenberg, A. M., Witt, J. K., & Metcalfe, J. (2013). From the revolution to embodiment: 25 years of cognitive psychology. *Perspectives on Psychological Science*, *8*(5), 573–585. doi:10.1177/1745691613498098 PMID:26173215

Goldin-Meadow, S. (2011). Learning through gesture. *Wiley Interdisciplinary Reviews: Cognitive Science*, *2*(6), 595–607. doi:10.1002/wcs.132 PMID:24187604

Goldin-Meadow, S. (2014). Widening the lens: What the manual modality reveals about learning, language and cognition. *Philosophical Transactions of the Royal Society, Biological Sciences.*, *369*(20130295).

Hanus, M., & Fox, J. (2015). Assessing the effects of gamification in the classroom: A longitudinal study on intrinsic motivation, social comparison, satisfaction, effort, and academic performance. *Computers & Education*, *80*, 152–161. doi:10.1016/j.compedu.2014.08.019

Hauk, O., Johnsrude, I., & Pulvermüller, F. (2004). Somatotopic representation of action words in human motor and premotor cortex. *Neuron*, *41*(2), 301–307. doi:10.1016/S0896-6273(03)00838-9 PMID:14741110

Johnson-Glenberg, M. C., Birchfield, D., Koziupa, T., & Tolentino, L. (2014). Collaborative Embodied Learning in Mixed Reality Motion-Capture Environments: Two Science Studies. *Journal of Educational Psychology*, *106*(1), 86–104. doi:10.1037/a0034008

Johnson-Glenberg, M. C., Birchfield, D., Megowan-Romanowicz, M. C., Tolentino, L., & Martinez, C. (2009). Embodied Games, Next Gen Interfaces, and Assessment of High School Physics. *International Journal of Learning and Media*, *1*(2). doi:10.1162/ijlm.2009.0017

Johnson-Glenberg, M. C., Birchfield, D., Megowan-Romanowicz, M. C., & Uysal, S. (2009). SMALLab: Virtual geology studies using embodied learning with motion, sound, and graphics. *Educational Media International*, *46*(4), 267–280. doi:10.1080/09523980903387555

Johnson-Glenberg, M. C., & Hekler, E. B. (2013). Alien Health Game: An Embodied, Motion-Capture Exer-Game Teaching Nutrition and MyPlate. *Games for Health Journal*, *6*(2). doi:10.1089/g4h.2013.0057

Johnson-Glenberg, M. C., Savio-Ramos, C., & Henry, H. (2014). ‘‘Alien Health’’: A Nutrition Instruction Exergame Using the Kinect Sensor. *Games for Health Journal: Research, Development, and Clinical Applications*, *3*(4), 241–251. doi:10.1089/g4h.2013.0094 PMID:25083315

Lindgren, R., & Johnson-Glenberg, M. C. (2013). Emboldened by Embodiment: Six Precepts regarding the Future of Embodied Learning and Mixed Reality Technologies. *Educational Researcher*, *42*(8), 445–452. doi:10.3102/0013189X13511661

Lindgren, R., & Moshell, J. M. (2011). *Supporting children's learning with body-based metaphors in a mixed reality environment*. Paper presented at the Proceedings of the 10th International Conference on Interaction Design and Children, Ann Arbor, Michigan. doi:10.1145/1999030.1999055

Lui, M., Kuhn, A., Acosta, A., Niño-Soto, M. I., Quintana, C., & Slotta, J. D. (2014). Using mobile tools in immersive environments to support science inquiry. *CHI '14 Extended Abstracts on Human Factors in Computing Systems 978*, 403-406. doi: 10.1145/2559206.2574796

Marlow, C. M. (2012). Making Games for Environmental Design Education: Revealing Landscape Architecture. *International Journal of Gaming and Computer-Mediated Simulations*, *4*(2), 60–84. doi:10.4018/jgcms.2012040104

Metz, K. (1985). The development of children's problem solving in a gears task: A problem space perspective. *Cognitive Science*, *9*(4), 431–471. doi:10.1207/s15516709cog0904_4

Milgram, P., & Kishino, A. F. (1994). Taxonomy of mixed reality visual displays. *IEICE Transactions on Information and Systems*, *E77-D*(12), 1321–1329.

Miller, A. (2012). Video Game Portal Enter the Classroom. Retrieved from http://blogs.kqed.org/mindshift/2012/07/video-game-portal-enters-the-classroom/

Mislevy, R. J., Behrens, J. T., DiCerbo, K. E., & Levy, R. (2012). Design and discovery in educational assessment: Evidence-centered design, psychometrics, and educational data mining. *JEDM-Journal of Educational Data Mining*, *4*(1), 11–48.

Nathan, M. J., Walkington, C., Boncoddo, R., Pier, E. L., Williams, C. C., & Alibali, M. W. (2014). Actions speak louder with words: The roles of action and pedagogical language for grounding mathematical reasoning. *Learning and Instruction, 33*, 182-193. doi: 10.1016

Papastergiou, M. (2009). Digital Game-Based Learning in high school Computer Science education: Impact on educational effectiveness and student motivation. *Computers & Education*, *52*(1), 1–12. doi:10.1016/j.compedu.2008.06.004

Segal, A., Black, J., & Tversky, B. (2010). *Do Gestural Interfaces Promotoe Learning? Congruent Gestures Promote Performance in Math*. Paper presented at the 51st Meeting of the Psychonomic Society Conference, St. Louis, Missouri.

Shaffer, D., Hatfield, D., Svarovsky, G., Nach, P., Nulty, A., Bagley, E., & Mislevy, R. et al. (2009). Epistemic Network Analysis: A Prototype for 21st Century Assessment of Learning. *International Journal of Learning and Media*, *1*(1).

Snow, E. L., Allen, L. K., Jacovina, M. E., & McNamara, D. S. (2015). Does agency matter?: Exploring the impact of controlled behaviors within a game-based environment. *Computers & Education*, *26*, 378–392. doi:10.1016/j.compedu.2014.12.011

Squire, K., & Klopfer, E. (2007). Augmented Reality Simulations on Handheld Computers. *Journal of the Learning Sciences*, *16*(3), 371–413. doi:10.1080/10508400701413435

Stafford, T., & Dewar, M. (2013). Tracing the trajectory of skill learning with a very large sample of online game players. *Psychological Science, online Dec 30*. doi: 10.1177/095679761351146

Vygotsky, L. S. (1978). *Mind in Society: The Development of Higher Psychological Processes*. Cambridge, MA: The Harvard University Press.

Wilson, M. (2003). Six views of embodied cognition. *Psychonomic Bulletin & Review*, *9*(4), 625–636. doi:10.3758/BF03196322 PMID:12613670

Young, M. F., Slota, S., Cutter, A. B., Jalette, G., Mullin, G., Lai, B., & Yukhymenko, M. et al. (2012). Our Princess Is in Another Castle: A Review of Trends in Serious Gaming for Education. *Review of Educational Research*, *82*(1), 61–89. doi:10.3102/0034654312436980

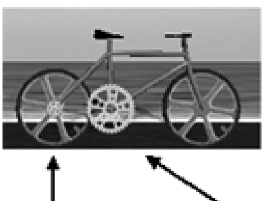
APPENDIX

Gear Test ONLY

Name:
Date:
Class Period:
Teacher's name:

Appendix Part 1.

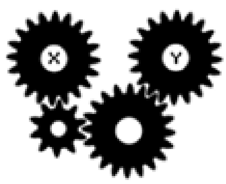
1. What is the gear ratio on these two bike gears?



Output gear = 10, Input gear = 50

2. If we triple to size of the input gear pictured above how much force will be required to go same speed?

a. more
b. less
c. the same amount as before



Use this picture to answer questions 3 and 4.

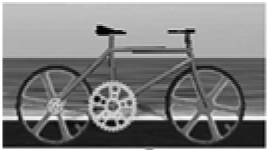
3. If gear X in the picture above turns clockwise (turns to the right) at a constant speed, how does gear Y turn?
   a. Clockwise (right) same speed
   b. Counterclockwise (left) same speed
   c. Clockwise (right) faster
   d. Counterclockwise (left) faster

4. When gear X has made 4 complete turns, how many turns will gear Y have made?
   a. 1
   b. 4
   c. 8
   d. 16

Appendix Part 2.

5. Explain how you arrived at this answer.

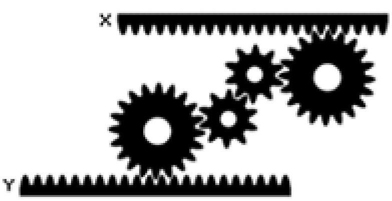
Use the picture below to answer questions 6 and 7, gears sizes are below.



Output Gear=5 Input Gear=7

6. Which completes one revolution first when a person is riding a bicycle?
   a. The pedal
   b. The rear wheel
   c. Both complete one revolution at the same time
   d. It varies depending upon what gear you are in

7. The Mechanical Advantage of the bicycle gear system pictured above is?
   a. 1.40
   b. .71
   c. 5.70
   d. 7.50



Use the image above to answer 8 and 9.

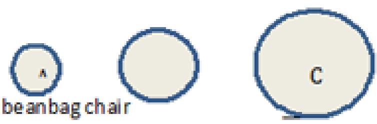
8. If Bar Y moves at a constant speed, how does bar X move?
   a. Faster
   b. Slower
   c. At the same speed
   d. It does not move

Appendix Part 3.

9. Assume there is an input gear to which a force is applied (this sets the gear train in motion), and an output gear that applies this force to the load – like a box that has to be lifted. If the input gear is 8 centimeters in diameter and the output gear is 16 centimeters in diameter, what is the gear ratio for this system?

a. 1:2
b. 2:1
c. 8:128
d. 128:8

10. This is an example of a gear system that is lifting an object with a long chain. The gear on the top is called input gear, you can crank that. The one on the bottom is called the output gear. A gear system like this will enable you to lift things that you are not otherwise physically capable of lifting. Imagine you need to move stuff into your second story bedroom. You have three gears you can choose from to use as input gears. Which of the three gears below would you choose as the **INPUT gear** to lift the following items? Write on the line next to the item the gear (A, B or C) you would choose to lift it.



i. A beanbag chair ____
ii. A candle ____
iii. A television ____
iv. A queen-size bed ____